MINING STRUCTURE PATTERNS BASED ON 3D FEATURES IN THE PROTEIN-DNA AND PROTEIN-PROTEIN COMPLEX

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Qing Sun

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

November 2018

Fargo, North Dakota

North Dakota State University
Graduate School

Title

MINING STRUCTURE PATTERNS BASED ON 3D FEATURES IN THE PROTEIN-DNA AND PROTEIN-PROTEIN COMPLEX

By

Qing Sun

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Changhui Yan
Chair

Anne Denton

Jun Kong

Xiwen Cai

Approved:

11/14/2018
Date

Kendall Nygard
Department Chair

ABSTRACT

For a long time, researchers have been searching for the "recognition codes" of protein complexes, which determine what DNA sequence or other protein a protein can bind to. The binding part prediction of protein complexes have important applications in many biological research fields, such as cleavage enzyme design and drug design and so on. Most sequence-based and PWM methods only capture sequence features on the protein interfaces and ignore the crucial spatial attributes of the features. This study investigates the recognition codes from a new angle, in which the preferred binding modes are captured using local structural motifs spanning the protein-DNA, protein-protein and protein-ligand interfaces. Using product graph, we transformed the structural motif discovery problem into a search for maximal cliques. These motifs include more information than the traditional amino acid-base contacting pairs. For example, in the protein-DNA interfaces research, we studied two domains, Zinc-finger and Helix-Turn-Helix (HTH), that both used a recognition helix to interact with DNA. In each domain, we found a few frequent structural motifs spanning the protein-DNA interfaces. Each motif includes at least 2 amino acids and 1 nucleotide from both sides of the interfaces. The motifs specify not only the types of amino acids and nucleotides involved in the interaction, but also the distances between them and their relative orientation.

The same method has been implemented in protein-protein and protein-ligand complexes. These motifs reveal preferred binding modes at the interfaces that involve more entities than the traditional contacting pairs. The biological and statistical significance of the motifs were confirmed using evolutionary conservation analysis and bootstrapping We also performed many other tests to evaluate our motifs' critical roles in the interactions. For example, we compared our motifs with experimentally verified hotspots. We also compared our method with other computational prediction methods to assess the effectiveness of the method. Our results

confirmed that the graph motifs discovered in this study play important roles in protein-DNA, protein-protein and protein-ligand interactions. We believe that the proposed graph method will be a very helpful tool for studying protein complexes interaction and other types of molecular interactions.

ACKNOWLEDGEMENTS

Firstly, I would like to express the deepest appreciation to my major advisor, Dr. Changhui Yan, who has the attitude and the substance of a respectable mentor: he continually and convincingly conveyed a spirit of adventure in research, and an excitement in regard to teaching. Without his guidance and persistent help, this dissertation would not have been possible.

I would also thank members of my doctoral committee: Dr. Anne Denton, Dr. Jun Kong and Dr. Xiwen Cai for their continued support and guidance.

I would like to thank my lab members, Wen Cheng, Liren Sun, Yongchao Ma and Yang Du for their friendship and collaboration. In our lab, someone was always able to lend an ear and provide friendly advice on matters both academic and personal.

I greatly appreciate the Department of Computer Science, North Dakota State University (NDSU) for providing me with this great opportunity to pursue my Ph.D. I am extremely grateful for all the friends and colleagues I have met during my doctoral studies.

Finally, I want to express my heartfelt thanks to my family: my parents, my wife Mingyi Zhang and my daughter Chloe. They are the best part of my life. Their love and encouragement inspire me to strive for ever greater heights.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

1. INTRODUCTION

With the almost entirely complete sequence of the human reference genome and numerous other genomes sequencing work, increasing availability of genome sequence data has led to the rapid growth of protein encoded information. Amino acids and nucleic acids are material basis of lives, and in the biological cell, proteins interact with each other and other biomolecules, such as DNA, to carry out specific functions.

One urgent task in the post-genomic era is to glean knowledge from this big data to elucidate various important biological processes. Protein-DNA interaction plays crucial roles in gene regulation, DNA transcription, replication, repair, and recombination. Therefore, protein-DNA interactions have been the subject of tremendous research efford in the past decades.

The high-resolution 3-dimensional structures of protein-DNA complexes show atomic details of the protein-DNA interfaces. Analyzing these complex structures, if available, can reveal the chemical and physical forces that facilitate the interactions. However, this kind of structures are very difficult to obtained using X-ray crystallography [4] and nuclear magnetic resonance (NMR) [5, 6] methods. Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) [7] is widely used to detect protein-DNA interactions in large scale. However, this method doesn't provide information regarding how the proteins and DNA bind. Due to the limits and difficulties associated with these experimental methods, computational methods have become an increasingly attractive approach to studying protein-DNA interaction.

Compared to traditional experimental methods, computational approaches can rapidly and accurately find the DNA-binding proteins under the premise of large data. According to the different aspects, sequence or structure, used in the computational methods, many famous and

stable machine learning methods and docking algorithms are applied in locating binding proteins with DNA in the complex structure.

2. LITERATURE REVIEW

2.1. Experimental Method

Protein-DNA interactions play fundamental roles in many biological processes, such as DNA replication, transcription, splicing, gene regulation, sequence encoding, and protein synthesis. The more protein-DNA complexes' structure we identify the better we will understand the mechanism of these vital processes in biological research. However, studies on protein-DNA recognition present complex technical challenges duo to the macromolecular structure. Therefore, the identification of structural features of protein-DNA complexes is at the cutting edge of protein-DNA interactome research.

To date, more than 5500 protein nucleic acid complex structures are searchable in Protein Data Bank (PDB) database which is significantly fewer than the actual number in nature. In the previous studies, distinguishing DNA bind proteins or their binding sites, finding the enough numbers of nucleic acid and protein sequences that exist have become the major aim of structural biology. Although a number of experimental methods such as electrophoretic mobility shift assays (EMSAs) [1, 2], conventional chromatin immunoprecipitation (ChIP) [3], MicroChIP [4], Fast ChIP [5], peptide nucleic acid (PNA)-assisted identification of RNA binding proteins (RBPs) (PAIR) [6], X-ray crystallography [7], and nuclear Magnetic resonance (NMR) spectroscopy [8], have been implemented in the structural mapping of protein-DNA complexes. However, these experimental methods are costly, time-consuming, labor-intensive or combination thereof.

2.2. Computational Method

With the development of bioinformatics, researchers have developed many computational approaches to predict the DNA binding proteins that are also suitable for RBPs (RNA binding proteins) prediction. Compared with the experimental approaches, computational methods could

identify DNA binding sites and RBPs rapidly and cheaply. During the past decades, vast quantities of genome sequences has been discovered with the development of the second-generation sequencing method (e.g., Illumine). This produces huge numbers of protein-DNA complexes and provides adequate data for the prediction of DNA-binding proteins and examining how interactions occur.

At this time, approaches can be divided into four categories based on the input features they used: sequence-based DNA-binding site prediction, structure-based DNA-binding site prediction, protein-DNA docking method and homology modeling and threading. Firstly, the sequence-based method applies the similarity of sequences to the identification between query sequences and sequences containing DNA-binding sites. Several studies have implemented this kind of method [9-13]. Although this sequences-based method can reach a rapid result, their performance is not satisfied enough. Both protein and DNA have complex spatial structures that cannot be represented by the sequence features. Duo to this defect, the structure-based methods were developed by researchers. The query protein-DNA complexes with unknown binding sites can be predicted by comparing with the known binding site structures [14-17]. It is believed that the structural similarity could provide more reliable and in-depth prediction consequence. Sequence similarity and structure similarity-based strategies are shown as follow:

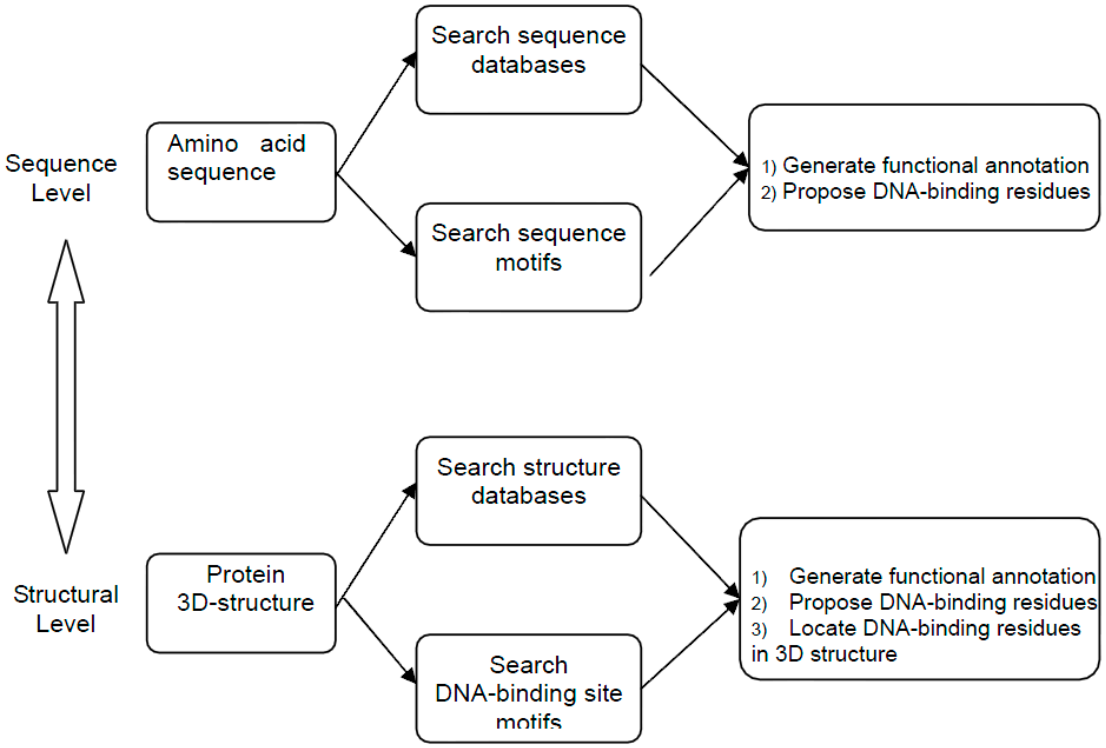

Figure 2.1. Sequence similarity and structure similarity-based strategies.

The protein-DNA docking method can help to modeling structures of a protein-DNA complex and to understand the mechanisms of interaction. This method is also frequently used to predict RBPs. For example, The Haddock [18] program which enables various molecules such as nucleotides, protein and other small molecules for docking. katchalski-katzir et al. [19] proposed a low-resolution docking program, which assigns different scoring functions for different ligands. Ritchie et al. [20] present a Hex method, which can perform protein-nucleotide docking and protein-protein docking. This program will give a unified score for all ligands and not special function for protein-DNA complexes. The FTDOCK program which is used by Gabb et al. [21] can accept both protein-protein docking and nucleotide-protein docking. Recently, two methods called “QUASI-RNP” and “DARS-RNP” have been developed by Tuszynsk and Bujnicki [22] to grade protein-RNA decoys using statistical and quasi-chemical reference.

Further, other specific residue-base interactions, for example particular hydrogen bond or structurally homologous proteins, can dock on DNA with relative spatial orientation [23-27].

The homology modeling and threading aim to model a query protein with an unknown structure by identifying a template protein of known structure. This method can generate a space model [28-30] by homology modeling or threading in the worst situation with no template protein to use. The accuracy of these situations will be weakened. In general, the homology modeling and threading method will be used as a complementary tactics, for example this method can used in meta-prediction [31] and comparative studies [32] to help increasing accuracy of the prediction of DNA-binding sites.

2.3. Recognition Code

Structural and biochemical studies of zinc finger proteins, for example Cys_2His_2 (C2H2), initially lead to the prediction of various "recognition code" which contains amino acids in the zinc finger protein and the bases in the corresponding DNA site. The ability to predict "recognition code" of protein or to effectively design the patterns that can bind to and release from their desired DNA target sites, could be extremely useful in many areas of biology and medicine [33].

Traci M Tanaka Hall [34] present a study of the versatile of zinc finger domains, they claimed that the C_2H_2 zinc finger scaffold play important roles on the prediction of base-specific recognition of the DNA major groove, backbone recognition of the RNA major groove, and almost customized RNA base and loop recognition. Yen Choo and Aaron Klug [35] gave the affirmative answer to whether a stereo chemical recognition code exists, which relates protein primary structure to DNA-sequence preference. They find that the contacts feature in the recognition code which shows that the binding mode of zinc fingers favors 1:1 amino acid and

base contacts in the plane of the base-pairs by observing the crystal structures of Zif268 and Ttk zinc finger complexes.

Jeffrey C. Miller and Carl O. Pabo [36] perform study at zinc finger variant which highlights another source of complexity involved in protein-DNA recognitions. They proposed that mutations could cause the rearrangement of key side-chains at the protein-DNA interface; these rearrangements would not result in fundamental changes in the spatial relationship between the polypeptide backbone and the DNA. Limin Angela Liu and Philip Bradley [37] claimed the importance of conformational flexibility for design and template-based modeling of protein-DNA binding complexes, and in this case, non-native conformations need to be sampled and accurately scored. They performed continued improvements in energy functions, solvation models and conformational sampling and they concluded that conformational flexibility increases prediction accuracy in the general situation when modeling interactions not directly visualized by the input structure.

Baldwin, E.P et al. [38] proposed the structural comparison of two Cre re-combinase variants in complex with different DNA sequences, and their results revealed that both DNA and protein differences affect the contacts made in the binding interface. For the purpose of account for the sequence degeneracy in transcription factor (TF)-DNA binding, the concept of position weight matrices (PWMs) was proposed and remains the most widely used representation of TF binding specificity [39,40]. PWMS are common representation scores for each DNA base pair along a binding site. PWM models can be learned from a variety of data types, from small collections of known binding sites to large datasets generated using high-throughput (HT) technologies [41]. PWMs can also be used to visualize as DNA logos, providing an intuitive feel for the TF-binding specificity [42]. HT studies are increasingly revealing unknown diversity in

DNA-binding preferences of numerous proteins, many of which may be the results of different binding modes [43-46].

Whereas it is impossible to generate a code which could be applicable to DNA-binding proteins of all known structural families. The structures of various complexes structural families show significant differences in the way the recognition helices of DNA-binding protein interact with the major groove of DNA. In contrast, it is still possible to predict the "recognition code" for complexes of a single structural family [47] or for a group of families which interact in similar ways with DNA [48].

2.4. Data-driven Approach

During the twentieth century, biology has finished the transition from a traditional descriptive science to a hypothesis-driven experimentation. This process is promoted by the increasing dominance of reductionism. In the general hypothesis-driven process, straight after researchers collect enough data, they need formulate a hypothesis about the aspects they are interested in. Then they perform experiments or observation to verify whether the hypothesis is correct or not. This method implicit that experiments and observations should only be made to support, or attack hypothesized mechanisms. Nowadays, researchers can find massive amounts of data benefiting from the development of information technology. They are more interested in the internal relation of data. In other words, researchers are attracted by the things they neither knew nor expected. So, with the availability of large datasets and advanced statistical and machine learning methods, many researchers presented their doubt whether we still need to rely on hypotheses in scientific inquiry [49]. Anderson [50] proposed that in the era of data explosion, the traditional hypothesis-driven scientific method would become obsolete. Other researcher [51] concluded that the hypotheses method is eventually limited because setting a premise before experiments would constrain them by established ways of thinking or doing.

To deal with the data deluge, the data-driven scientific method emerged. Researchers [52] claim this method with no more theories or hypotheses, no specific experimental results to refute or support the hypotheses. Equipped with such huge data sets, we can perform data mining in an objective way. The first step of data-driven method is to identify the specific data you are really interested in. Then implement a data analysis approach to measure or describe attributes of the data. The critical section of data-driven process are the sophisticated algorithms and statistical tools used here. Finally, try to extract crucial relation between data and your experimental results. In this sense, the data-driven approach can be seen as a hypothesis generator, not a hypothesis tester. The goal is to discover correlations and connections between properties of huge size data and to dig new things we neither knew nor expected [53]. In this way, data-driven method is very helpful to extract and convert implicit data information into new knowledge by apply algorithms mining data for plausible patterns [54]. Also, it's a novel approach for scientific researchers increasing the possibilities of breakthrough areas where nobody had looked before.

In this research we performed a data-driven approach to extract crucial patterns from protein complex dataset. All processes in our research are compelled by data, rather than by intuition or personal experience. We collected our protein complex data from many famous databases containing experiment verified data. So, the data we used are accessible, queryable and trustworthy. We also implemented strategies for the data cleaning. This step aims to remove structures with insufficient information and structures with high similarity. Data cleaning also guarantee our data credible and representative. Then we implemented innovative graph-based methods to extract knowledge and insights from our data to get crucial patterns for protein complexes. The validity of these patterns was subsequently assessed by various statistical modeling programs. The comparison of the results of this novel approach with analyses by other

computational methods supports the claim that this data-driven approach is capable of identifying biologically relevant patterns and associations.

3. OBJECTIVES

1) The main objective of this project is to propose a graph method for the discovery of structure patterns in the protein-DNA interfaces. The discovered patterns will help understand how proteins interact with DNA to achieve desired binding affinity and specificity.
2) The project aims to perform an extension study: verify the roles coevolution residues and conservations played in the DNA-binding site prediction.
3) The third objective of this project is to modify the proposed method and make it applicable for other type of prediction, for example the Zinc-Finger and Helix-to-Helix prediction of residues in the protein-DNA interaction.
4) The last objective of this project is to apply the proposed method to other type of interactions, such as protein-protein interaction and protein-ligand interaction. Verify the validity of our method.

4. EXPERIMENTAL APPROACH

4.1. Mining of Common Sub-graphs at DNA-binding Sites

In this work, we present a graph method for the discovery of structure patterns in the protein-DNA interfaces. The discovered patterns will help understand how proteins interact with DNA to achieve desired binding affinity and specificity.

4.1.1. Datasets

We extracted all protein-DNA complexes from the PDB [55]. Then, the dataset was culled based on protein sequence similarity using PISCES [56]. The resulting dataset (will be referred to as Dataset I) consisted of 308 proteins-DNA complexes with mutual sequence identity ≤ 30% and each protein had at least 40 amino acid residues. All the structures have resolution better than 3.0 Å and R factor less than 0.3. This dataset will be used to discover common patterns enriched in the DNA-binding sites. Dataset II was the van Dijk and Bonvin protein-DNA docking benchmark [57]. It consisted of both bound and unbound structures of 47 protein-DNA complexes, ranked from easy to difficult according to how much the conformation change during binding. The unbound structures will be used to generate docking poses for the complexes. Then, different scoring methods are used to assign scores to the conformations. The performance of the scoring functions will be evaluated by comparing the docking conformations with the native bound structures.

4.1.2. Interface Residues

Interface residues on the DNA-binding sites were defined as in [58]. We used NACCESS [59] software to calculate the accessible surface area (ASA) of each amino acid in both bound and unbound states. An amino acid was defined as an interface residue if its ASA in unbound state was at least 1Å2 more than that in bound state.

4.1.3. Graph Representation of DNA-binding Sites

Each DNA-binding site was represented using a graph, where each node represented an interface residue and an edge was added between two nodes if the corresponding residues were in contact. Two residues were considered contacting if the nearest distance between their heavy atoms was less than 0.5Å. Each node was labeled with its residue type. Each edge was also associated with an edge label. If the two nodes at both ends of an edge were sequence neighbors on the protein chain, then the edge was labeled as type one; otherwise, the edge was labeled as type two.

4.1.4. Mining of Common Sub-graphs

There were 308 protein-DNA complexes, which were encoded as 308 graphs using the graph representation mentioned above. We implemented the VF2 algorithm [60] to discover common sub-graphs between each pair of graphs. In the test of isomorphism, we took into consideration the node labels and edge labels. In this study, we focused on the common sub-graphs that had at least 3 nodes, because, common sub-graphs with less than 3 nodes contain too few information.

4.2. Evaluate the Common Sub-graphs

4.2.1. Discovery of Graph Patterns Enriched in the Protein-DNA interfaces

First, 308 DNA-binding sites were extracted from Dataset I and represented as a graph. These graphs will be referred to as binding-site graphs. The VF2 algorithm was used to find common sub-graphs between each pair of binding-site graphs. After removing duplicated common sub-graphs, we obtained 24,356 unique common sub-graphs. In order to find the graph patterns that occurred with higher frequencies in the DNA-binding sites than in other regions of the protein surface, we randomly collected 308 non-binding sites from the 308 proteins, with one non-binding site from each protein. The non-binding site from a protein had the same size as the

DNA-binding site from the same protein and there was no overlap between the non-binding site and DNA-binding site. These non-binding sites served as the background control for the identification of common sub-graphs enriched in the DNA-binding sites. The non-binding sites were also represented as graphs.

For each sub-graph, we checked whether it occurred in the 308 binding-site graphs and the 308 non-binding site graphs. The presence or absence of a sub-graph in the binding-site and non-binding sites was recorded using a vector of 616 values, with 1 being presence and 0 absences. Then, we performed a t-test to identify sub-graphs that enriched in the DNA-binding sites. A lower p value given by the t-test indicated that the sub-graph was more favored in the DNA-binding sites. At the end, we obtained 2,594 sub-graphs with p values less than 0.05. Among them, 600 had 3 nodes, 1,349 had 4 nodes, and 645 had five or more nodes. These are the graph patterns that had higher propensities to occur in the DNA-binding sites than in other regions of the proteins.

4.2.2. Discrimination between DNA-Binding Sites and Non-DNA-Binding Sites

To evaluate the significance of the discovered graph patterns, we used the patterns as features to discriminate DNA-binding sites from non-DNA-binding sites. When n patterns were used, a binding site or non-binding site was encoded as an n-value vector, with values 1 and 0 denoting the presence and absence, respectively, of a graph pattern on the site. The 308 DNA-binding sites and 308 non-DNA-binding sties were used to train and test classifiers using 10-fold cross validation. We tried five popular classification methods implemented in Weka [61], including Random Forest, Support Vector Machine (SMO), Random Committee, Bayesian, and J48. Among these methods, the SMO (with RBF kernel) achieved the best results. For the SMO method, we also tested different number of patterns as encoding features. The 2,594 sub-graph patterns were sorted in the order of increasing p value. In each experiment, n patterns from the

top of the list were used to encode the protein patches. N varied from 100 to 2,500 with increments of 100. Our results showed that the accuracy of the classification increased as n increased, arriving at the best accuracy when 2,200 patterns were used. After that, the accuracy decreased when n continued to increase. When the 2,200 patterns were used, the SMO discriminated DNA-binding sites versus non-DNA-binding sites with 79.1% accuracy, 88.3% specificity, and 69.8% sensitivity. This result suggested that the sub-graphs discovered in the above section revealed the structural patterns that facilitated the interaction between protein and DNA and could be used to predict DNA-binding sites on protein structure.

4.2.3. Overlap between Graph Patterns and UniProtKB Annotations

To further evaluate the biological significance of the discovered graph patterns, we compared them with the annotations in UniProtKB [62], a comprehensive database of protein functional information. In a UniProtKB entry, we searched for fields that provided information regarding which amino acid residues were involved in DNA binding. We found three fields that contained such information: the REGION subsection denoted the stretch of protein sequence that matched a function domain; the SITE subsection described interesting single amino acid sites on the sequence; and the MUTAGEN subsection described the effect of experimental mutation of one or more amino acids on the biological properties of the protein. The REGION subsection usually contained a long stretch of protein sequence and did not provided information regarding specific amino acids. Thus, we focused on the SITE and MUTAGEN fields. Specifically, for the 308 proteins in Datatset I, we searched the MUTAGEN subsection to look for amino acids whose mutation resulted in decrease of DNA-binding affinity, and we also searched the SITE subsection to look for amino acids that were involved in DNA-binding. We compared these residues with the residues on the proteins that were covered by the discovered sub-graph patterns. We were able to find DNA-binding residues in the SITE subsection in 21 proteins, and

in 16 of them these residues overlapped with the residues that were covered by the sub-graph patterns (Table I). We were also able to find DNA-binding residues in the MUTAGEN subsection of 39 proteins, and in 20 of them these residues overlapped with the sub-graph patterns (Table II). These results confirm the biological significance of the sub-graph patterns. We noticed that in a few proteins, there was no overlap between UniProtKB annotations and the sub-graph patterns. One possible explanation for this is that the information gathered from mutagenesis experiments (MUTGEN) and SITE is scarce. Thus, they only cover a small fraction of DNA-binding residues.

Table 4.1. Overlap between subgraph patterns and UniProtKB site annotations

PDBID	# of residues covered by sub-graph patterns	# of residues found in the UniProtKB SITE subsection	# of residues overlapped
1MW8X	24	8	6
1A31A	26	9	5
3MVAO	64	5	5
1I7DA	37	5	3
1DC1A	29	2	2
1MUSA	37	2	2
2EX5A	42	4	2
3H0DA	22	2	2
3MLNA	34	2	2
1OZJA	14	1	1
1RFFA	18	1	1
1XJVA	28	1	1
2VY1A	8	4	1
3AAFA	19	2	1
3FDEA	31	1	1
4GLXA	75	2	1
1MNNA	30	6	0
2W36A	25	1	0
2XCSD	34	2	0
3GNAA	9	1	0
3IGKA	5	1	0

Table 4.2. Overlap between subgraph patterns and UniProtKB mutagen annotations

PDBID	# of residues covered by sub-graph patterns	# of residues found in the UniProtKB MUTAGEN subsection	# of residue overlapped
2BZFA	12	8	4
3KDEC	9	5	4
3B39A	8	3	3
3RNUA	4	8	3
1FLOA	44	2	2
1OZJA	14	2	2
1RFFA	18	6	2
1CL8A	12	1	1
1H9DAB	18	10	1
1J3EA	23	8	1
1K3XA	19	2	1
1T9IA	33	8	1
2C7PA	45	1	1
2FMPA	48	4	1
2PY5A	40	3	1
2W36A	25	1	1
3AAFA	19	3	1
3MLNA	34	5	1
3O1TA	6	5	1
3QE9Y	23	3	1
1JEYAB	31	4	0

Table 4.2. Overlap between subgraph patterns and UniProtKB mutagen annotations (continued)

PDBID	# of residues covered by sub-graph patterns	# of residues found in the UniProtKB MUTAGEN subsection	# of residue overlapped
1MNNA	30	6	0
1ZRFA	17	16	0
2BOPA	10	3	0
2BSQAF	7	1	0
2C62A	12	8	0
2EX5A	42	1	0
2O4AA	13	8	0
2R9LA	21	4	0
2VY1A	8	4	0
3COQA	5	1	0
3G9MA	17	1	0
3GNAA	9	5	0
3H0DA	22	1	0
3H15A	7	2	0
3MVAO	64	6	0
3QMDA	16	2	0
3U4QAB	32	3	0
4ECQA	40	2	0

4.2.4. A Scoring Function for Protein-DNA Docking based on Sub-graph Patterns

When molecular docking is used to predict the structure of a protein-DNA complex, the unbound structures of the protein and the DNA are used as input to generate a large number of possible poses that the protein-DNA complex may take. Poses that are similar to the native

structure of the protein-DNA complex are usually known as near native poses and poses that are not similar to the native structure are called docking decoys. Then a scoring function is used to assign scores to these poses. A good scoring function should assign higher scores to near native poses than docking decoys. In previous sections, we have discovered a set of patterns that are favored in the protein-DNA interfaces. In this section, we will test the patterns' ability to discriminate near native protein-DNA poses from docking decoys. For this purpose, we built a simple scoring function that counted the number of sub-graph patterns that occurred in the protein- DNA interface of a pose and assign that number as a score to the pose. The rationale of this design is that since these patterns are favored in the protein-DNA interfaces, a near native pose is more likely to have these patterns to occur in the protein-RNA interface than a docking decoy. We will compare our scoring method with the scoring method used in FTdock [63], a well-established docking method based on shape complementarity, electrostatics, and biochemical information.

In a previous work, [64] used FTdock to generate 100,000 poses for each complex in Dataset II. They also ranked the generated poses using the scoring method implemented in FTdock. For the sake of reducing computational time, we used the 1,000 best poses for each protein-DNA complex from [64]. We assessed the performance of a scoring method using the same procedure proposed by [64]. For each protein-DNA complex, the root means square deviation (RMSD) values between the 1,000 poses and the native complex structure were calculated, and the 20 poses that had the lowest RMSD values were identified and they were named 20 best RMSD poses. Then, the scoring method was used to rank the 1,000 poses. A prediction set was populated by gradually recruiting more and more poses from the top of the ranking. The scoring function was assessed by measuring the fraction of the 20 best RMSD

poses that appeared in the prediction set as the size of the prediction set increased. A good scoring function should put the 20 best RMSD poses on the top of the ranking and thus, the fraction should approach 1 quickly as the size of the prediction set increases.

Figure 4.1 shows the comparison between FTdock scoring function and our proposed scoring function. When the size of the prediction set was less than 75, the FTdock method included higher fraction of the 20 best RMSD poses in the prediction set. When the prediction size was 75, both methods included 10 of the 20 best RMSD poses in the prediction set. When the size continued to increase, the fraction achieved by our proposed method was higher and approached 1 faster than FTdock. The area under curve (AUC) is 0.85 for our method and o.75 for FTdock. This result suggests that while the two methods complement each other at two ends of the spectrum. Our method has slightly better overall performance. We also examined the rank of the best RMSD pose that had the lowest RMSD value. On average, our scoring function ranked the best RMSD pose on the 102nd position while FTdock put it on the 124th position. This result showed that our scoring method is slightly better than FTdock in identifying the best pose.

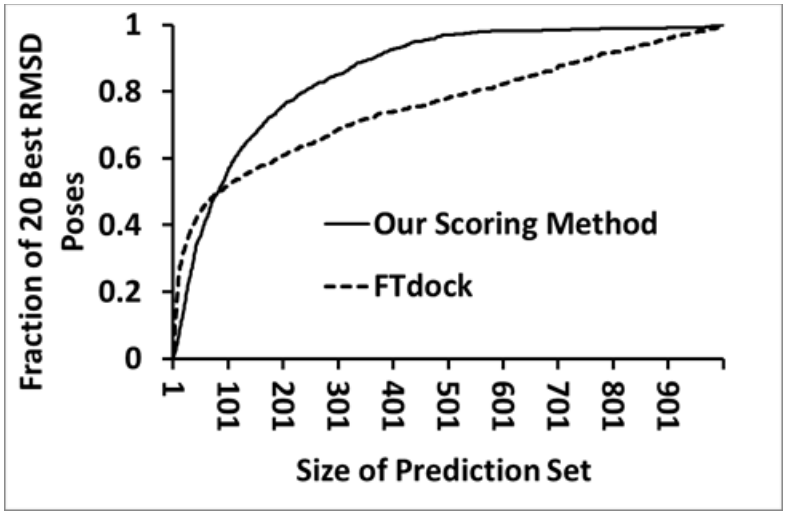

Figure 4.1. Comparison between our scoring method and the scoring method of FTdock

4.2.5. Conclusions

In this part, we used an innovative graph representation method to encode protein structure and discovered structure patterns that were favored in the protein-DNA interfaces. These patterns may play critical roles in the protein-DNA interaction, while further work is still needed to elucidate how they contribute to the binding affinity and specificity. To evaluate the significance of these patterns, we used them as features to build classifiers to discriminate DNA-binding sites versus non-DNA-binding sites. The SMO classifier was able to predict DNA-binding sites with 79.1% accuracy, 88.3% specificity, and 69.8% sensitivity. The biological significance of the patterns was further confirmed by comparing them with expert knowledge in UniProtKB. The comparison showed that the discovered patterns had significant overlap with amino acids sites that were considered crucial for DNA-binding in UniProtKB. We also demonstrated that the patterns could be used to build a simple scoring method to discriminate near native docking poses from docking decoys. The proposed method was much simpler than FTdock scoring method and yet achieved slightly better performance. Our results confirmed that the graph patterns discovered in this study play important roles in protein-DNA interactions, and the pattern mining method proposed in this study will be a very useful tool for the investigation of interactions between biological macromolecules.

4.3. Conservation and Coevolution Calculation

4.3.1. Calculation of Mutual Information

We analysis the original 308 PDB files, get the amino acids sequences of them. Then we search each sequences in the “Treefam” database and download the alignment files with “.fa” format if there is a successful search result. We finally get 128 alignment result files among the 308 sequences.

Next step we calculated the mutual information using alignment files. We calculate the mutual information between every two locations in the PDB sequences base on its multiple sequence alignment (MSA) files. For example, there are two locations X and Y in the "1A31" amino acids sequence, the method to calculate their MI between them is as follow:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Where p(x, y) is the joint probability distribution function of X and Y, P(X) and P(Y) are the marginal probability distribution functions of X and Y respectively. After calculation we get 128 result files record the mutual information scores of the 128 alignment files. We also normalized the mutual information scores for each file and sort the MI scores.

4.3.2. Construct Co-evolution Network

In this section we construct two kinds of contact network using the spatial coordinates and co-evolution scores of amino acids.

A. Construct 3D contact network

The 3D contact network of all amino acids and surface amino acids will be used to edge filter of the co-evolution network. We calculate the 3D contact network using the spatial coordinates (can be obtained in the PDB files) of the amino acids in a sequence. In order to contrast in the next part, we need contrast two kinds of 3D contact network, one contains all the amino acids and the other one just contains the surface amino acids.

In order to represent amino acids using networks we need calculate the relationship of each two nodes. Like the traditional network, our representation network is constituted by points and edges connect points. In the network each amino acid is represented by a node and we put an edge between two nodes if the corresponding residues were considered connected. Then we calculate the nearest distance between the heavy atoms from two residues so that we can decide

whether they are defined connected or not, if the two amino acids are sequence neighbors or the nearest distance between them is less than 0.5Å, we put an edge into the representation network between these two nodes.

B. Filter Co-evolution network

In this part we will first construct the original Co-evolution network using the sorted MI scores of the 128 sequences. We must set a threshold value to filter the amino acid and the edges among them. We construct the original Co-evolution networking using the amino acids with higher MI scores for than the threshold value. For example, we set the threshold value as 0.5 and we will put an edge between two nodes (amino acids) if the MI score between them is higher than 0.5.

After getting the original co-evolution network of each alignment files of PDB sequences we need filter them using the 3D contact network. We will get the whole nodes' co-evolution network filtered by the 3D contact network of all amino acids and the surface nodes' co-evolution network filtered by the 3D contact network of surface amino acids. When the original co-evolution network are filtered by the 3D contact network of all amino acids, we retain all the amino acids as nodes and retain all the edges appear in both original co-evolution network and the 3D contact network of all amino acids; When the original co-evolution network are filtered by the 3D contact network of surface amino acids, we retain all the surface amino acids as nodes and retain all the edges appear in both original co-evolution network and the 3D contact network of surface amino acids. The two kinds of co-evolution network will be used in the next comparative experiment.

4.3.3. Total Contact vs Co-evolution Contact

In this part we contrast the contact number of 3D contact network and co-evolution network in “Whole Nodes” and “Surface Nodes” ways. We set three different threshold values 0.5, 1, 2 respectively to perform the horizontal and vertical comparison. Also, we identify the DNA-binding amino acids using the NACCESS software to help result analysis. The contrast results are show as table4.3-4.8:

Table 4.3. Total contact vs co-evolution contact in whole nodes and MI>0.5

Whole Nodes	**Non-DNA-binding**	**DNA-binding**
Total # of 3D co-evolution contact	498329	50769
Total # of 3D contact	648465	76767
Percentage	0.7685	0.6613

Table 4.4. Total contact vs co-evolution contact in surface nodes and MI>0.5

Surface Nodes	**Non-DNA-binding**	**DNA-binding**
Total # of 3D co-evolution contact	275758	37102
Total # of 3D contact	343100	55442
Percentage	0.8037	0.6692

Table 4.5. Total contact vs co-evolution contact in whole nodes and MI>1

Whole Nodes	**Non-DNA-binding**	**DNA-binding**
Total # of 3D co-evolution contact	428724	41687
Total # of 3D contact	648465	76767
Percentage	0.6611	0.5430

Table 4.6. Total contact vs co-evolution contact in surface nodes and MI>1

Surface Nodes	**Non-DNA-binding**	**DNA-binding**
Total # of 3D co-evolution contact	246553	30832
Total # of 3D contact	343100	55442
Percentage	0.7186	0.5561

Table 4.7. Total contact vs co-evolution contact in whole nodes and MI>2

Whole Nodes	**Non-DNA-binding**	**DNA-binding**
Total # of 3D co-evolution contact	258579	21579
Total # of 3D contact	648465	76767
Percentage	0.3988	0.2811

Table 4.8. Total contact vs co-evolution contact in surface nodes and MI>2

Surface Nodes	**Non-DNA-binding**	**DNA-binding**
Total # of 3D co-evolution contact	160635	15958
Total # of 3D contact	343100	55442
Percentage	0.4682	0.2878

When contrast in the horizontal dimension, take the threshold is 0.5 as example, as table 1 shows the percentage of 3D contact number divided by 3D co-evolution contact is about 0.77 in the "Non DNA-binding" nodes part which is bigger than the 0.77 in the "DNA-binding" nodes part; it is same case in the table 2 of surface condition when 0.80 of "Non DNA-binding" nodes part is bigger than the 0.67 in the "DNA-binding" nodes part. When we raise the threshold to 1 and 2, we get the same result as threshold=0.5. This means that the co-evolution amino acids may not contribute so much to the DNA-binding sites prediction.

When contrast in the vertical dimension, we take the "Whole Nodes" tables of threshold is 0.5, 1 and 2 to analysis firstly, the percentage of 3D contact number divided by 3D co-evolution contact are 0.77, 0.66 and 0.40 in the "Non-DNA-binding" nodes part, and 0.66, 0.54 and 0.28 in the "DNA-binding" nodes part. Then for the "Surface Nodes" tables of threshold are 0.5, 1 and 2, the percentage of 3D contact number divided by 3D co-evolution contact are 0.80, 0.72 and 0.47 in the "Non-DNA-binding" nodes part, and 0.67, 0.56 and 0.29 in the "DNA-binding" nodes part. As the threshold increase, the MI scores of nodes are increase too, but the

percentage of “Non-DNA-binding” and “DNA-binding” both decrease, which means that the co-evolution feature has little effect in DNA-binding sites prediction.

4.3.4. Degree of Intermediate and Non-intermediate

In this part we contrast the average degree of “DNA-binding” nodes and “Non-DNA-binding” nodes in whole nodes’ co-evolution network and surface nodes’ co-evolution network. We still set three different threshold values 0.5, 1, 2 respectively to perform the horizontal and vertical comparison. The contrast results are show as table4.9-4.14:

Table 4.9. Degree of intermediate and non-intermediate in whole nodes and MI>0.5

Whole Nodes	**DNA-binding**	**Non-DNA-binding**
Total Degree	16403	181941
Total # of node	3919	32174
Average	4.1855	5.6549
Rate	0.7401	
# of 0-degree nodes	2152	13979

Table 4.10. Degree of intermediate and non-intermediate in surface nodes and MI>0.5

Surface Nodes	**DNA-binding**	**Non-DNA-binding**
Total Degree	13063	123197
Total # of node	3776	23510
Average	3.4595	5.2402
Rate	0.6602	
# of 0-degree nodes	2100	9167

Table 4.11. Degree of intermediate and non-intermediate in whole nodes and MI>1

Whole Nodes	**DNA-binding**	**Non-DNA-binding**
Total Degree	9411	111801
Total # of node	3919	32174
Average	2.4014	3.4749
Rate	0.6911	
# of 0-degree nodes	2572	17332

Table 4.12. Degree of intermediate and non-intermediate in surface nodes and MI>1

Surface Nodes	DNA-binding	Non-DNA-binding
Total Degree	7541	79479
Total # of node	3776	23510
Average	1.9971	3.3806
Rate	0.5907	
# of 0-degree nodes	2448	11424

Table 4.13. Degree of intermediate and non-intermediate in whole nodes and MI>2

Whole Nodes	DNA-binding	Non-DNA-binding
Total Degree	2244	31572
Total # of node	3919	32174
Average	0.5726	0.9813
Rate	0.5835	
# of 0-degree nodes	3348	24717

Table 4.14. Degree of intermediate and non-intermediate in surface nodes and MI>2

Surface Nodes	DNA-binding	Non-DNA-binding
Total Degree	1780	23522
Total # of node	3776	23510
Average	0.4714	1.0005
Rate	0.4712	
# of 0-degree nodes	3136	16703

When contrast in the horizontal dimension, take the threshold is 0.5 as example, as table 7 shows the average degree of “DNA-binding” nodes is about 4.19 which is smaller than the 5.65 in the “Non DNA-binding” nodes part; it is same case in the table 8 of surface condition when 3.46 of “DNA-binding” nodes part is smaller than the 5.24 in the “Non DNA-binding” nodes part. The percentage of “Non-DNA-binding” divided by “DNA-binding” in “Whole Nodes” co-evolution network is 0.74 which is bigger than 0.66 in “Surface Nodes” co-evolution network. When we raise the threshold to 1 and 2, we get the same result as threshold=0.5. This means that the co-evolution amino acids may not contribute so much to the DNA-binding sites prediction.

When contrast in the vertical dimension, we take the “Whole Nodes” co-evolution network tables of threshold is 0.5, 1 and 2 to analysis firstly, the average degree is 4.19, 2.40 and

0.57 in the “DNA-binding” nodes part, and 5.66, 3.47 and 0.98 in the “Non-DNA-binding” nodes part. Then for the “Surface Nodes” co-evolution tables of threshold are 0.5, 1 and 2, the average degree is 3.46, 2.00 and 0.47 in the “DNA-binding” nodes part, and 5.24, 3.38 and 1.00 in the “Non DNA-binding” nodes part. As the threshold increase, the MI scores of nodes are increase, but the average degree and the percentage of “Non-DNA-binding” divided by “DNA-binding” in “Whole Nodes” co-evolution network and “Surface Nodes” co-evolution network are all decrease, which means that the co-evolution feature has little effect in DNA-binding sites prediction.

4.3.5. Conservation and Coevolution

In this part, we use three different methods “Consurf”, “Rate4” and “Shannon” to calculate the conservation of each amino acid of the 128-sequence using the alignment files of them. We also draw the coordinate diagram in which conservation scores indicating X-axis and coevolution scores indicating Y-axis for each method. It is worth mentioning that the smaller the “Consurf” conservation score is the more conservative it is; it is same for the “Rate4” method and in contrast with the “Shannon” method. For the coevolution scores we choose the highest score for each amino acid. We also use different color to distinct DNA-binding nodes and Non-DNA-binding nodes in the diagram.

Figure 4.2 shows the result diagram of “Consurf” results.

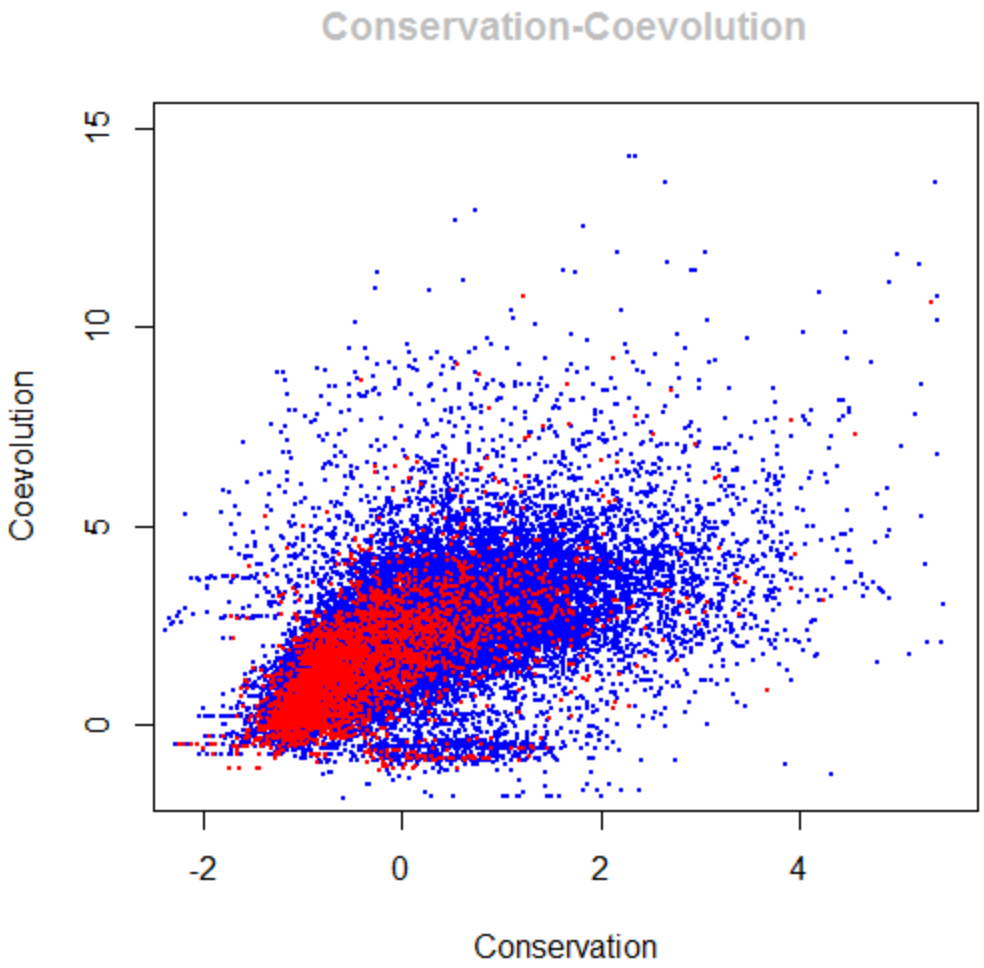

Figure 4.2. Largest coevolution of MI (Consurf)

In the Figure 4.2 the red points represent DNA-binding nodes and the blue points represent Non-DNA-binding nodes. We can see from the diagram that the DNA-binding nodes concentrated in the lower left corner of the diagram which means that the node with lower coevolution scores (lower coevolution property) and lower conservation scores (higher conservation property) is more likely to be the DNA-binding sites.

Figure 4.3 shows the result diagram of “Rate4” results.

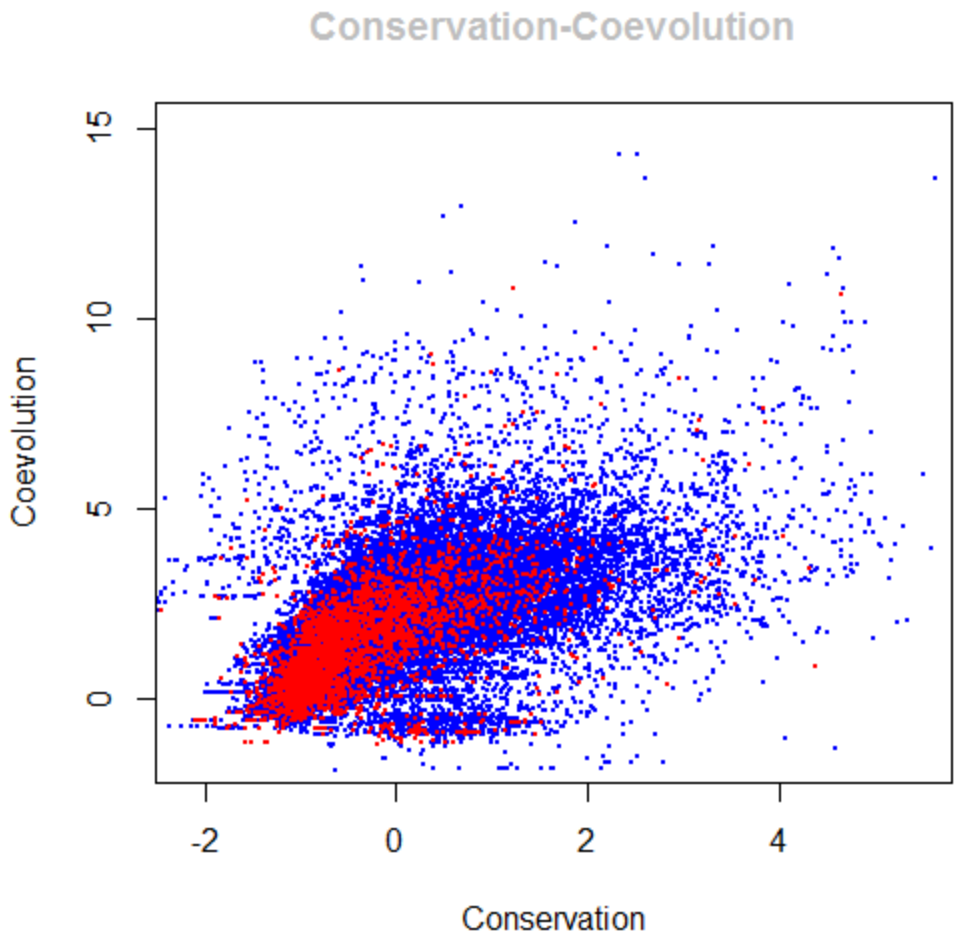

Figure 4.3. Largest coevolution of MI (rate4)

Figure 4.3 use the same method to show DNA-binding nodes and Non-DNA-binding nodes. We can see from the diagram that the node with lower coevolution scores (lower coevolution property) and lower conservation scores (higher conservation property) is more likely to be the DNA-binding sites, just like the Figure 4.2.

Figure 4.4 shows the result diagram of “Shannon” condition.

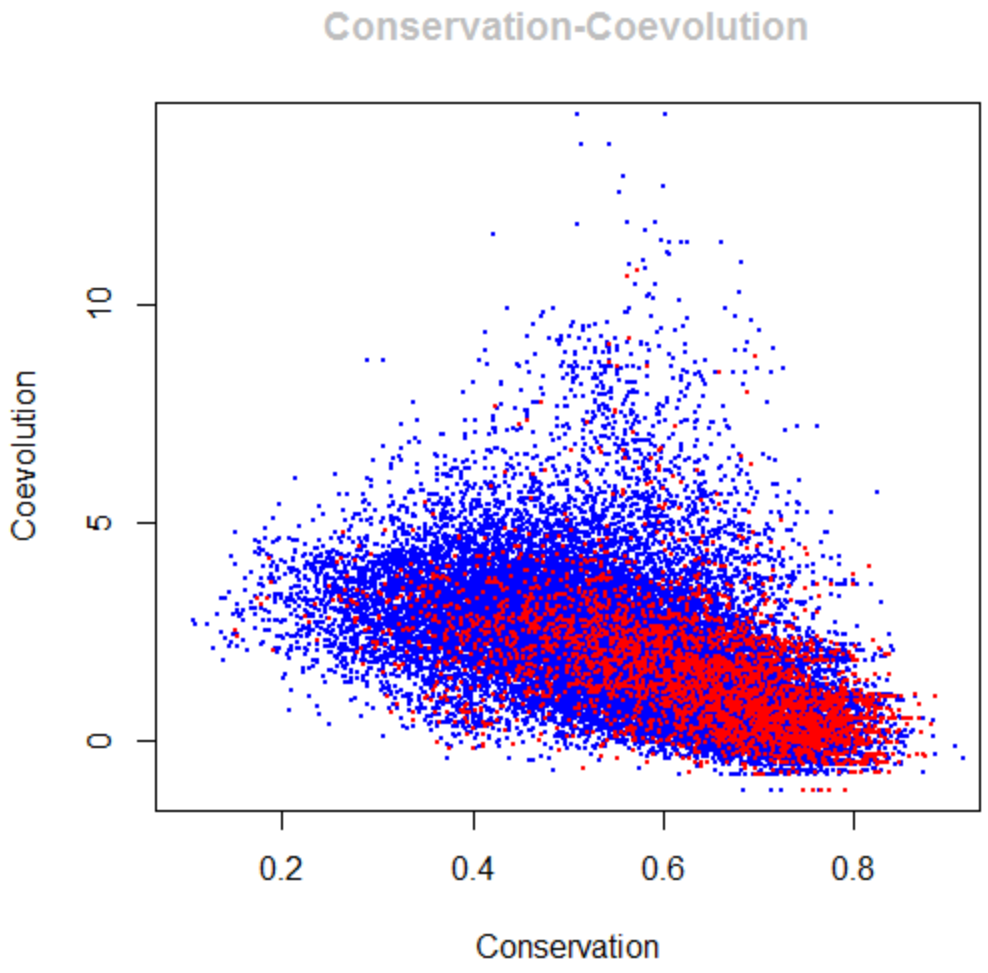

Figure 4.4. Largest coevolution of MI (Shannon)

We can see from Figure 4.4 that the DNA-binding nodes concentrated in the lower right corner of the diagram which means that the node with lower coevolution scores (lower coevolution property) and higher conservation scores (higher conservation property) is more likely to be the DNA-binding sites.

From the analysis of the three methods, we can conclude that the conservation property has more impact in DNA-binding site prediction than the coevolution property. In order to further confirm our conclusion, we also create the diagram with conservation scores indicating

X-axis and the average coevolution scores indicating Y-axis for “Rate4” method because this method has demonstrated our conclusion more obvious.

Figure 4.5 shows the new result diagram of “Rate4” condition.

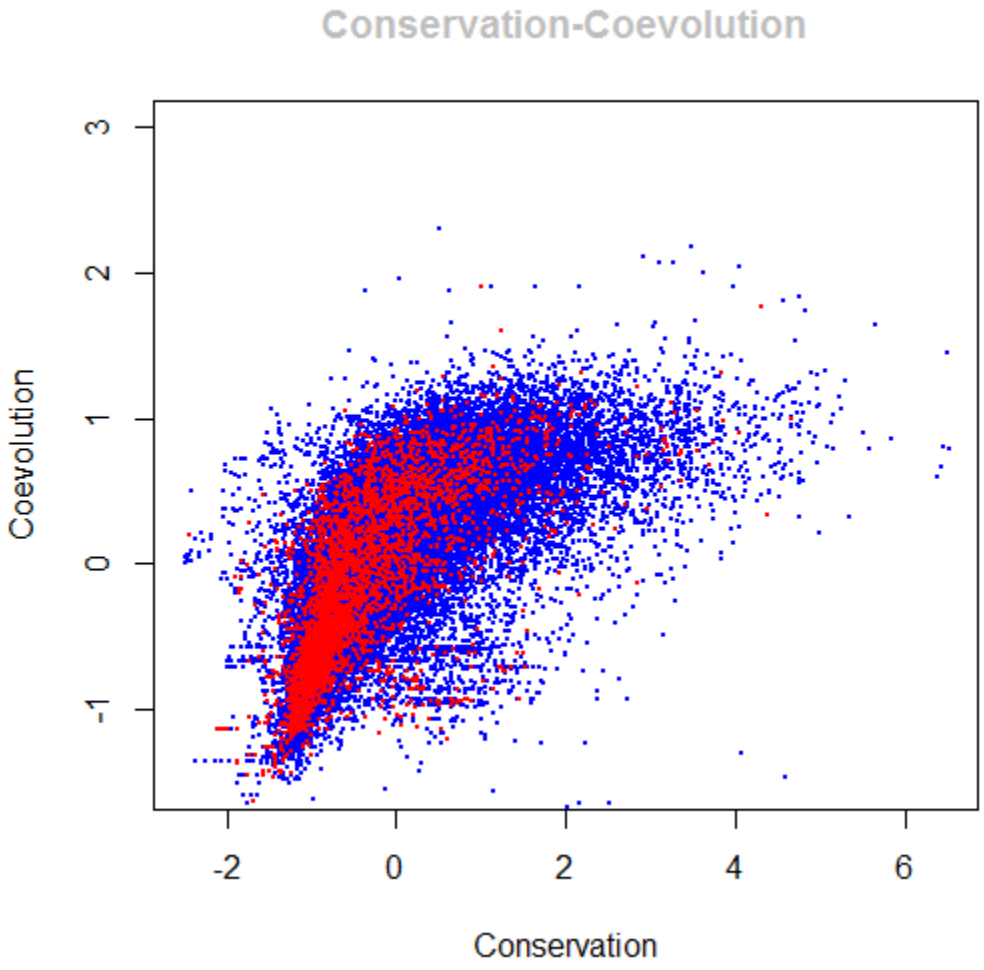

Figure 4.5. Average coevolution of MI (Rate4)

The Figure 4.5 shows that the DNA-binding nodes concentrated in the lower left corner of the diagram which still supports our previous conclusion that the conservation property is more important than coevolution property in DNA-binding site prediction.

4.3.6. Segment Contrast

In this part, we show the distribution of DNA-binding sites and Non DNA-binding sites in different segment of coevolution scores and conservation scores. Figure 6 shows the amino acid distribution with largest coevolution scores of sites.

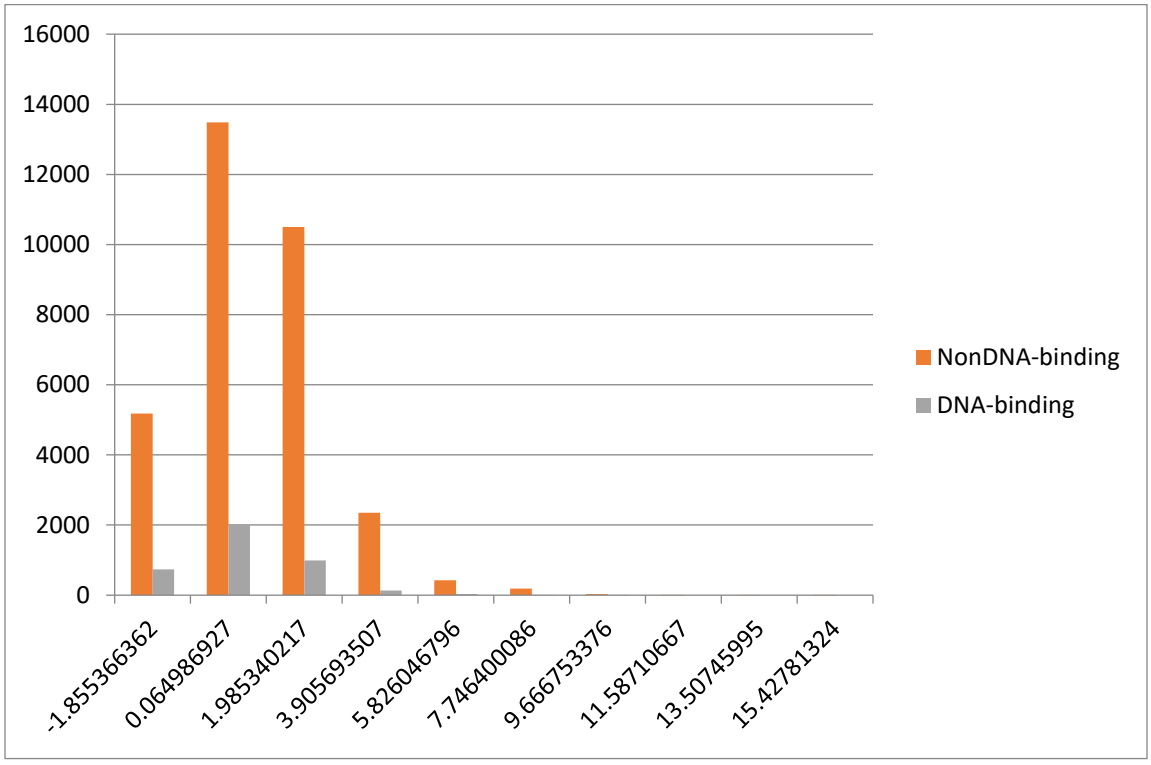

Figure 4.6. Largest coevolution scores of sites

We can see from the Figure 4.6 that almost all the DNA-binding sites concentrate on the lower coevolution area which means the coevolution property has little impact in DNA-binding site prediction.

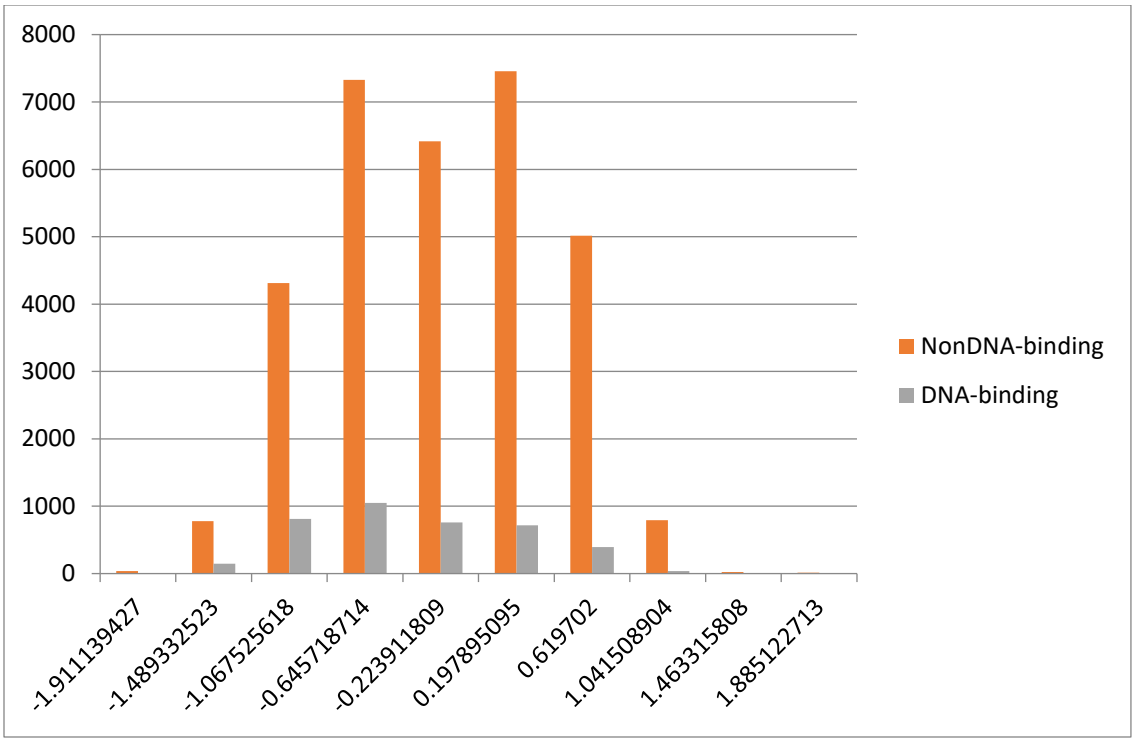

Figure 4.7. Average coevolution scores of sites

We can see from the Figure 4.7 that the DNA-binding sites average distribute on the segment of coevolution area which means the coevolution property has little impact in DNA-binding site prediction.

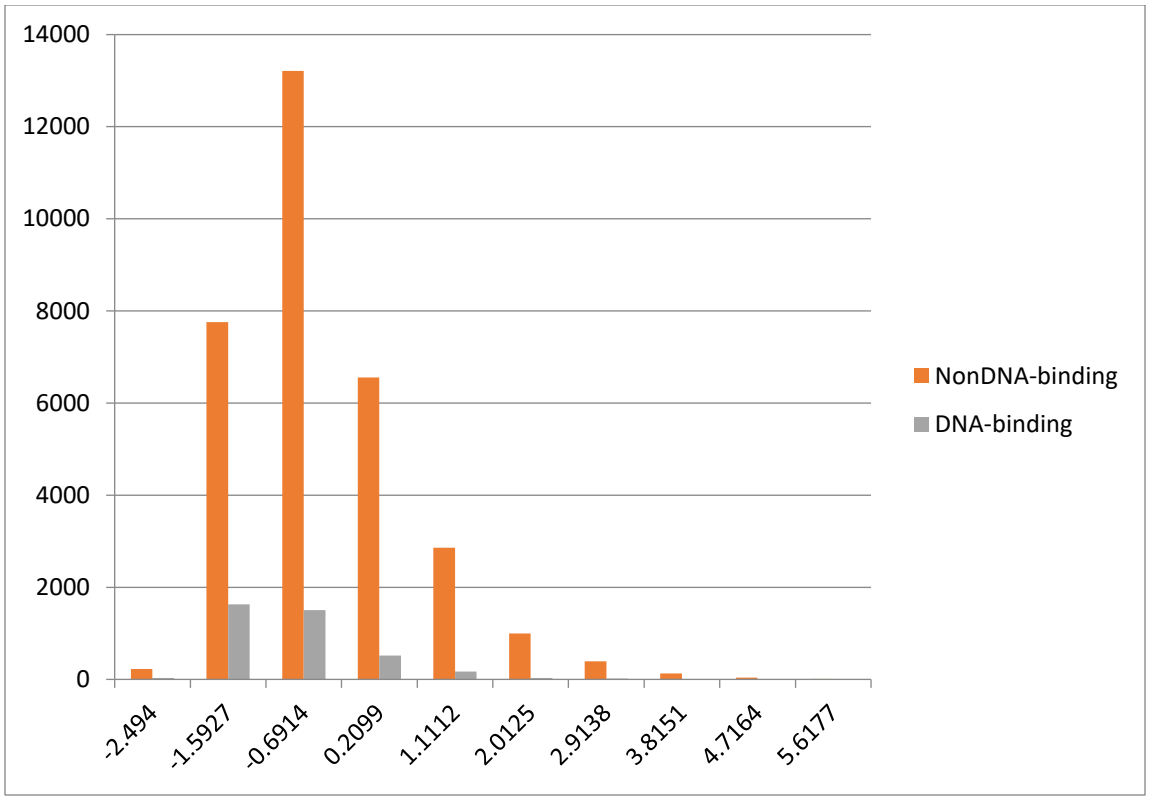

Figure 4.8. Conservation scores of sites (Rate4)

We can see from the Figure 4.8 that almost all the DNA-binding sites concentrate on the lower conservation area which means the conservation property has more impact in DNA-binding site prediction.

4.4. Apply the Proposed Method to Other Type of Interactions Part I

The interactions between DNA and protein can be formed by different domains, such as the zinc finger or the helix-turn-helix. In this part, we apply our new structural method for prediction of recognition motif in these two helix regions. We take both amino acids of protein and nucleotides of DNA into consideration when compare the different helix region from Protein-DNA complexes. We use the common motif to reveal the principles that rule the complexity of recognition code in DNA-binding pattern. In the prediction process, we also apply the 3D distance and relative position among amino acids and nucleotides to the construction of

helix graph. In the validation phase, we confirmed the statistical significance of our motifs by calculating the frequencies of occurrences of our motifs in various random distribution situations.

4.4.1. Dataset Preparation

We collected two kinds of datasets from Astral Scope 2.06 genetic domain sequence subsets, based on PDB SEQRES records, with less than 40% identity to each other. The type number of Zinc-Finger dataset is g.37.1.1 and the type number of Helix-to-Helix is a.4.1.1.

The original datasets contain 100 items of Zinc-Finger position and 21 items of Helix-to-Helix position. Duo to not all of these items can be used in our research; we performed some data Filtering work. Firstly, we filtered these items by PDB ids and reserved items involving both DNA chain and amino acid chains in their PDB structures, secondly find helix region in amino acid chain for each PDB structures through the database of "Protein Data Bank", "Structural Classification of Zinc Fingers" and "InterPro" website. The results are listed in Table 4.15 and Table 4.16:

Table 4.15. Zinc-Finger dataset

Zinc-Finger	Chain	Helix index region
2GLI	A	153-166,183-194,217-225,244-257
1LLM	C	115-128,143-155
1A1I	A	119-130,147-158,175-185
2DRP	A	122-134,152-164

Table 4.16. Helix-to-Helix dataset

Helix-to-Helix	Chain	Helix index region
3A01	A	225-245
1PUF	A	245-268
1E3O	C	141-157
4J19	A	322-344
1PUF	B	276-294
2HOS	A	41-60
1K61	A	172-189

There is maybe more than one helix region in a PDB structure and for every helix region we will build an original graph using amino acids and nucleotides as nodes.

4.4.2. Graph Construction

In this part we constructed 3D graphs combining nucleotides and amino acids by using the filtered data of the two datasets. Our method is that we chose one helix region and all the nucleotides in the same PDB structure to construct an original graph every time. For the Zinc Finger dataset and Helix-to-Helix dataset we will get 11 graphs and 7 graphs respectively. Every node of the graph represents a nucleotide or amino acid. In this step we divided 20 amino acids into 7 types based on the side chain classes (Wikipedia), the classification is as Table 4.17:

Table 4.17. Classification of amino acids

Side-chain class	Abbreviation	Amino Acid
Aliphatic	ALI	ALA, GLY, ILE, LEU, VAL
Basic	BAS	ARG, HIS, LYS
Acid/Amide	ACI	ASN, ASP, GLU, GLN
Sulfur-containing	SUL	CYS, MET
Aromatic	ARO	PHE, TRP, TYR
Cyclic	CYC	PRO
Hydroxyl-containing	HYD	SER, THR

We also take the 3D position of every molecular into consideration in the graphics building process, so we need calculate the distance between them. According to molecular contact properties, we use the side chain atoms of amino acids and the base atoms of nucleotides to calculate the distance and chose the minimum one as the edge length. For the edge length of graphs, we set two thresholds: 10 Å for the edge between same type of nodes (both nodes are amino acids or nucleotides) and 5 Å for the edge between different type of nodes (one node is amino acid and the other one is nucleotide). The remaining edges will help us to filter the nodes by removing the nodes with no edge connect to them. At last we get the final original helix graphs for both Zinc Finger and Helix-to-Helix datasets.

4.4.3. Calculation of Common Sub-graphs

During the construction of original helix graphs, we take both node type and 3D position into consideration and in this part of calculation of common sub-graphs we still consider these two aspects. For each dataset we calculate their common sub-graph of every three graphs using

maximal clique algorithm, the absolute difference of edge length between common Parton in these three graphs is less than or equal to 1 Å. We get 89 common sub-graphs (C0-C88) from Zinc-Finger dataset. The results are show in Table 4.18:

Table 4.18. Common sub-graphs of Zinc-Finger dataset

Type of vertices in the spatial motif	**Spatial motif**	**How many proteins have it?**	**How many times does it occur? (Which original graph does it occur?)**
(1) ALI, ACI, DG	C0	2	3(1,2,6)
(2) ACI, BAS, DG	C1, C2, C18, C26	4	14(1,2,3,4,5,6,8,9)
	C39, C40, C48, C49, C50, C52, C60, C61, C62, C64, C65, C70, C71, C72, C73, C74, C75, C78, C83, C88	4	19(1,2,4,5,6,7,8,9,10)
	C44, C56, C79	3	7(4,5,7,10)
(3) ACI, BAS, DC	C3, C4, C5, C10, C13, C14, C19, C41, C42, C43, C38, C47, C55, C59, C63, C76, C77	4	22(1,4,5,6,7,8,10)
	C57, C84, C87	3	8(4,5,7,9,10)
(4) ACI, HYD, DC	C6, C7, C9, C11, C12, C20, C21, C66, C68, C69	3	8(1,4,7,8,10)
(5) ACI, HYD, DC, DC	C8	3	3(1,4,7)
(6) ACI, ACI, DT	C15, C17, C22, C80	4	4(1,5,8,10)
(7) ACI, ACI, DC	C16	4	5(1,5,8,10)
(8) BAS, BAS, DG	C23, C35, C36, C37	4	12(2,3,4,5,6,9)

Table 4.18. Common sub-graphs of Zinc-Finger dataset (continued)

Type of vertices in the spatial motif	**Spatial motif**	**How many proteins have it?**	**How many times does it occur? (Which original graph does it occur?)**
(9) ACI, BAS, DT	C24	3	4(2,4,6,7)
	C53, C54	3	8(4,5,6,9,10)
(10) BAS, HYD, DC	C25	3	4(2,3,4,7)
	C27, C28	3	7(2,3,4,5,8)
(11) BAS, ALI, DT	C29	2	4(2,9,10)
(12) BAS, HYD, DT	C30	3	3(3,4,7)
	C31, C32, C33, C34	4	11(3,4,7,9)
(13) ACI, BAS, DC, DG	C45	3	6(4,5,7,9,10)
	C51	3	11(4,5,6,7,10)
	C67	3	7(4,5,7,10)
(14) ACI, BAS, DT, DG	C46	2	3(4,5,7)
(15) ACI, BAS, DA	C81, C82, C85, C86	2	5(5,9,10)

There are 15 categories of spatial motif in Zinc Finger original graphs and every motif contains one or more common sub-graphs. Among each type of motif, we range the same common sub-graphs (represent same part of original graphs) into a group by putting them in the same line of the table. The table also list the number of PDB structures containing the common sub-graph group, the number of occurrences of this group and which original graph contains it.

Among the 15 categories there are 6 categories having sub-groups. They are category 2, 3, 9, 10, 12 and 13. We also calculate the RMSD of side length between different sub-group motifs in one category. The results are shown in Table 4.19:

Table 4.19. RMSD of groups in Zinc-Finger common sub-graphs

Category	**Containing groups**	**RMSD between groups**
2	2, 3, 4	2:3=2.622063510353115 2:4=1.8521920783344679 3:4=1.184452043023833
3	5, 6	5:6=1.9176873665617455
9	12, 13	12:13=3.2634564564368387
10	14, 15	14:15=1.4271028182584606
12	17, 18	17:18=3.2111184925875262
13	19, 20, 21	19:20=1.0644598483078727 19:21=1.2924677455035813 20:21=0.9159951255934112

We can see from the table the RMSD between different groups in one category is around 1or greater than 1, this means there are significantly differences between these subgroups even though they have same points.

As shown in the Zinc-finger table, there are 23 groups or lines (from 1 to 23) in total. We calculated the internal RMSD of side length using the average graph of each group and the result is shown in Figure 4.9:

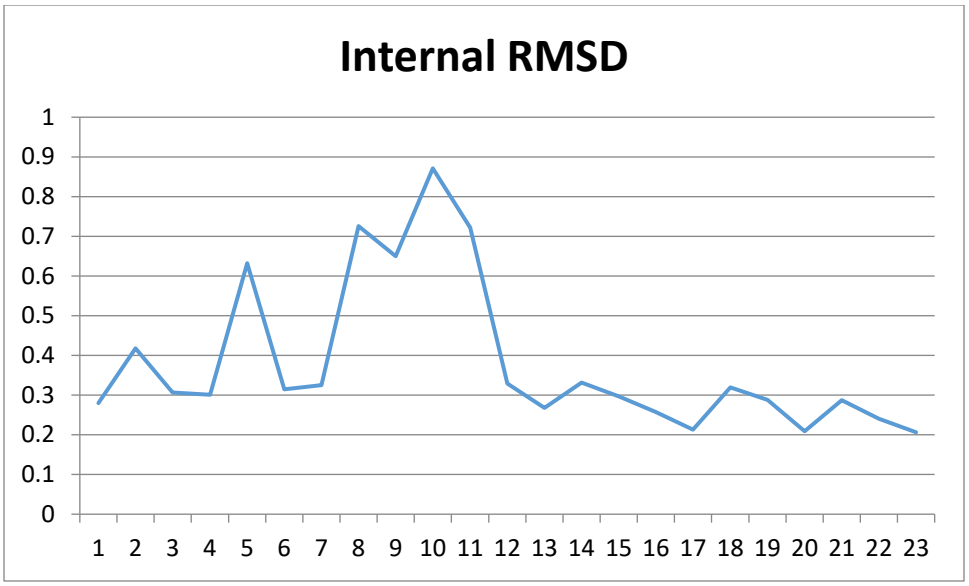

Figure 4.9. Internal RMSD of groups in Zinc-Finger common sub-graphs

We also get 53 common sub-graphs (C0-C52) from Helix-to-Helix dataset. The results are in Table 4.20:

Table 4.20. Common sub-graphs of Helix-to-Helix dataset

Type of vertices in the spatial motif	**Spatial motif**	**How many proteins have it?**	**How many times does it occur? (Which original graph does it occur?)**
(1) ACI, BAS, DG	C25	4	8(1,2,3,6)
(2) ACI, ACI, DC	C31	4	4(1,2,3,6)
(3) BAS, BAS, DG	C3	5	6(0,1,3,5,6)
(4) ACI, BAS, DT	C1, C4, C7, C16, C18, C22, C39, C51	6	13(0,1,2,3,5,6)
	C11, C14, C28, C32	6	12(0,1,2,3,4,5,6)
	C19, C33, C34, C36, C37	5	17(0,1,2,4,5,6)
(5) ARO, ACI, DA	C0, C13, C15, C17, C26, C27, C29, C30, C38, C45, C47, C48, C49, C50	5	12(0,1,3,4,5,6)
(6) ACI, ARO, ACI, DA	C2, C6, C20, C35	3	4(0,1,4,6)
	C46	3	3(3,4,6)
(7) BAS, BAS, DT	C5, C8, C9, C10, C23, C24	5	9(0,1,2,5,6)
	C52	5	5(2,3,4,5,6)
(8) ACI, HYD, DA	C12	3	7(0,2,6)
(9) ACI, ACI, DA	C21	3	5(0,4,6)
	C40, C41, C42	4	5(2,3,4,6)
(10) ACI, ALI, DA	C43, C44	4	6(2,3,4,5)

There are 10 categories of spatial motif in Helix-to-Helix original graphs and every motif contains one or more common sub-graphs. The table is organized using the same format with the Zinc Finger table.

Among the 10 categories there are 4 categories having sub-groups. They are category 4, 6, 7 and 9. We also calculate the RMSD of side length between different sub-group motifs in one category. The results are shown in Table 4.21:

Table 4.21. RMSD of groups in Helix-to-Helix common sub-graphs

Category	**Containing groups**	**RMSD between groups**
4	4, 5, 6	4:5=1.4247767560363612 4:6=2.038416936864183 5:6=2.288778979678851
6	8, 9	8:9=2.3718044921211425
7	10, 11	10:11=1.6604455476684572
9	13, 14	13:14=2.493068664546218

We can see from the table the RMSD between different groups in one category are all greater than 1, this means there are significantly differences between these subgroups even though they have same points.

As shown in the Helix-to-Helix table, there are 15 groups or lines (from 1 to 15) in total. We calculated the internal RMSD of side length using the average graph of each group and the result is shown in Figure 4.10:

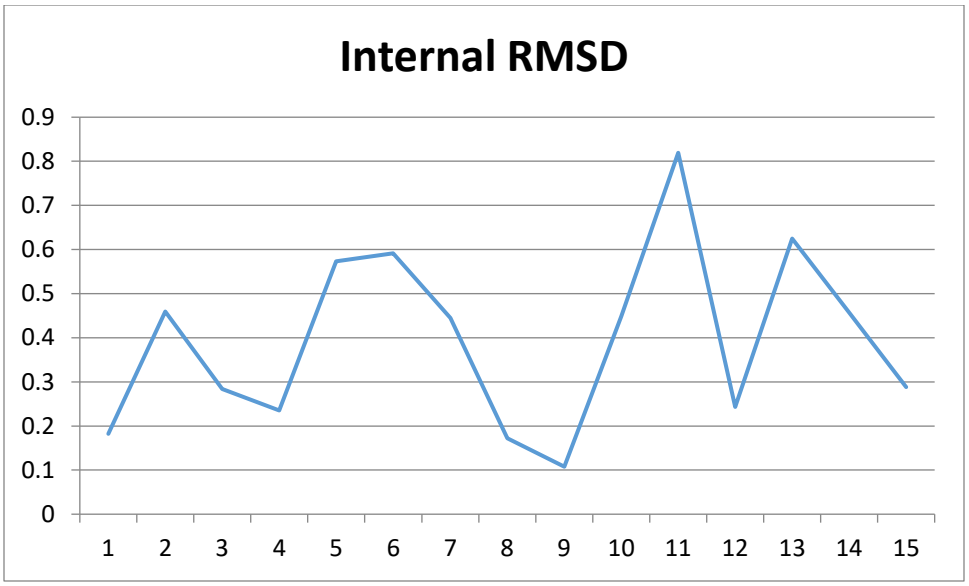

Figure 4.10. Internal RMSD of groups in Helix-to-Helix common sub-graphs

4.4.4. Analysis of Common Groups

From the tables of two datasets, we find 4 common type motifs between them accord to the vertices in the spatial motif. They are list in Table 4.22:

Table 4.22. Common motifs between Zinc-Finger and Helix-to-Helix common sub-graphs

Type of vertices in the spatial motif	Zinc Finger	Helix-to-Helix
ACI, BAS, DG	Motif (2)	Motif (1)
ACI, ACI, DC	Motif (7)	Motif (2)
BAS, BAS, DG	Motif (8)	Motif (3)
ACI, BAS, DT	Motif (9)	Motif (4)

By analyzing and comparing the edge distance of these pairs of motifs (absolute value of the corresponding edge length difference <=1) we find a common group between them. They are the “C24” group of Zinc Finger and the “C19, C33, C34, C36, C37” group of Helix-to-Helix. We also find corresponding sites of the PDB structure of this two groups and calculate the

conservation of these sites. For the protein chain we use “ConSurf” server to get the normalized conservation scores (average score for all residues is zero, and the standard deviation is one). The conservation scores calculated by ConSurf are a relative measure of evolutionary conservation at each sequence site of the target chain. The lowest score represents the most conserved position in a protein. We use “Jaspar” and “Transfac” (if have corresponding data) sever to calculate the conservation of nucleotide chain and give out the percentage of conservation. The results of the group “C24” of Zinc Finger dataset is shown in Table 4.23 and 4.24:

Table 4.23. Conservation of points in Zinc-Finger motifs

PDB	**ACI**	**Conservation**	**BAS**	**Conservation**	**DT**	**Conservation (Jaspar, Transfac)**
2GLI	ASN-A222	-0.206	ARG-A217	-0.425	C9	31.0%
1LLM	ASP-C117	-0.558	HIS-C122	-1.23	B30	6.671%,100%
1A1I	ASP-A121	1.657	ARG-A127	-0.566	C53	3.97%, 87.2%
	ASP-A148	-0.472	HIS-A153	-0.84	B5	94.0%, 83.6%

Table 4.24. Conservation of points in Helix-to-Helix motifs

PDB	ACI	Conservation	BAS	Conservation	DT	Conservation (Jaspar, Transfac)
3A01	ASN-A235	-0.762	ARG-A242	-0.446	C9	100%
	ASN-A235	-0.762	LYS-A230	-0.99	C12	57.9%
	ASN-A235	-0.762	ARG-A242	-0.446	D7	89.5%
1PUFA	ASN-A255	-0.952	LYS-A262	-0.930	D11	96.7%,92.9%
	ASN-A255	-0.952	LYS-A262	-0.930	E29	99.2%,92.9%
	ASN-A255	-0.952	LYS-A261	-0.690	E29	99.2%,92.9%
	ASN-A255	-0.952	LYS-A250	-0.844	E26	92.2%
	ASN-A255	-0.952	LYS-A261	-0.690	E26	92.2%
1E3O	ASN-C151	-0.977	ARG-C146	-0.734	A206	100%,100%
1PUFB	ASN-B286	-1.41	LYS-B293	-0.217	E33	94.4%,100%
	ASN-B282	-1.081	ARG-B288	-1.42	E29	88.9%,97.5%
2HOS	ASN-A51	-0.734	LYS-A58	-0.718	D31	100%
	ASN-A51	-0.734	LYS-A46	-0.661	C14	93.6%
	ASN-A51	-0.734	LYS-A58	-0.718	C11	97.3%
1K61	ASN-A178	-0.907	ARG-A183	-0.96	E15	93%
	ASN-A178	-0.907	ARG-A184	-0.96	F26	93.8%
	ASN-A178	-0.907	ARG-A185	-0.886	E15	93%

We can see from the table that the “ACI” and “BAS” sites’ conservation score are all below zero except “A121” in “1A1I”. It does not necessarily indicate 100% conservation (e.g. no mutations at all), but rather indicates that this position is the most conserved in this specific protein (DNA) calculated using a specific MSA. The “DT” sites’ conservation have one or more percentages, the first one is calculated by “Jaspar” matrix and the second one is calculated by “Transfac” matrix (if have) combining with some alignment method. All these two percentages are almost more than 50% and some reach 90% even 100%. The 31.0% of “2GLI” is the largest one among A C G and T, the 6.671% of “1LLM” and 3.97% of “1A1I” is the second largest one. All this effective proof that the corresponding sites in PDB structure of the common group between Zinc Finger sub-graph and Helix-to-Helix sub-graph is conservation and are functional sites with a great possibility.

4.4.5. Distribution Calculation

To further proof the statistics significant of this common group Parton, we perform the randomly distribution experiment for each dataset. We reconstruct the original graph by redefine the type nodes through the occupancy of this type of nodes. The more occupancy they have the larger probabilities they are assigned. The method helped to build new randomly designed original graphs for both datasets. Then we will calculate the total number of occurrences of our common group Parton in these new original graphs. We operate this process 100 times for both datasets and the results are shown in Table 4.25 and Table 4.26:

Table 4.25. Distribution of Zinc-Finger motifs

Type of vertices in the spatial motif	Spatial motif	Actual number of occurrences	Average of 100 trials of randomly distribution	Standard deviation of 100 trials of randomly distribution
(1) ALI, ACI, DG	C0	3	2	1.220
(2) ACI, BAS, DG	C1, C2, C18, C26	14	4.38	2.301
	C39, C40, C48, C49, C50, C52, C60, C61, C62, C64, C65, C70, C71, C72, C73, C74, C75, C78, C83, C88	19	2.52	1.472
	C44, C56, C79	7	2.05	1.412
(3) ACI, BAS, DC	C3, C4, C5, C10, C13, C14, C19, C41, C42, C43, C38, C47, C55, C59, C63, C76, C77	22	4.85	2.986
	C57, C84, C87	8	3.39	2.001
(4) ACI, HYD, DC	C6, C7, C9, C11, C12, C20, C21, C66, C68, C69	8	1.18	0.898
(5) ACI, HYD, DC, DC	C8	3	0.18	0.313
(6) ACI, ACI, DT	C15, C17, C22, C80	4	0.65	0.806
(7) ACI, ACI, DC	C16	5	1.31	1.402
(8) BAS, BAS, DG	C23, C35, C36, C37	12	4.29	2.420

Table 4.25. Distribution of Zinc-Finger motifs (continued)

Type of vertices in the spatial motif	Spatial motif	Actual number of occurrences	Average of 100 trials of randomly distribution	Standard deviation of 100 trials of randomly distribution
(9) ACI, BAS, DT	C24	4	0.77	0.886
	C53, C54	8	2.03	1.540
(10) BAS, HYD, DC	C25	4	2.91	1.539
	C27, C28	7	3.84	1.766
(11) BAS, ALI, DT	C29	4	3.49	2.139
(12) BAS, HYD, DT	C30	3	1.68	1.106
	C31, C32, C33, C34	11	6.74	2.470
(13) ACI, BAS, DC, DG	C45	6	0.7	0.840
	C51	11	0.56	0.750
	C67	7	0.63	0.819
(14) ACI, BAS, DT, DG	C46	3	0.37	0.570
(15) ACI, BAS, DA	C81, C82, C85, C86	5	0.74	0.844

Table 4.26. Distribution of Helix-to-Helix motifs

Type of vertices in the spatial motif	Spatial motif	Actual number of occurrences	Average of 100 trials of randomly distribution	Standard deviation of 100 trials of randomly distribution
(1) ACI, BAS, DG	C25	8	1.3	0.974
(2) ACI, ACI, DC	C31	4	1.16	1.117
(3) BAS, BAS, DG	C3	6	2.43	1.685
(4) ACI, BAS, DT	C1, C4, C7, C16, C18, C22, C39, C51	13	5.69	2.467
	C11, C14, C28, C32	12	7.82	3.416
	C19, C33, C34, C36, C37	17	2.44	1.370400
(5) ARO, ACI, DA	C0, C13, C15, C17, C26, C27, C29, C30, C38, C45, C47, C48, C49, C50	12	1.34	1.169
(6) ACI, ARO, ACI, DA	C2, C6, C20, C35	4	0.02	0.039
	C46	3	0.08	0.149
(7) BAS, BAS, DT	C5, C8, C9, C10, C23, C24	9	5.64	2.994
	C52	5	4.9	2.622
(8) ACI, HYD, DA	C12	7	1.65	1.275

Table 4.26. Distribution of Helix-to-Helix motifs (continued)

Type of vertices in the spatial motif	Spatial motif	Actual number of occurrences	Average of 100 trials of randomly distribution	Standard deviation of 100 trials of randomly distribution
(9) ACI, ACI, DA	C21	5	2.82	2.075
	C40, C41, C42	5	0.86	0.877
(10) ACI, ALI, DA	C43, C44	6	1.53	1.345

We do more detailed analysis of the common pattern of the two datasets: group "C24" of Zinc-finger dataset and group "C19, C33, C34, C36, C37" of Helix-to-Helix datasets. The result of 100 trials of randomly distribution of them is show in Figure 4.11:

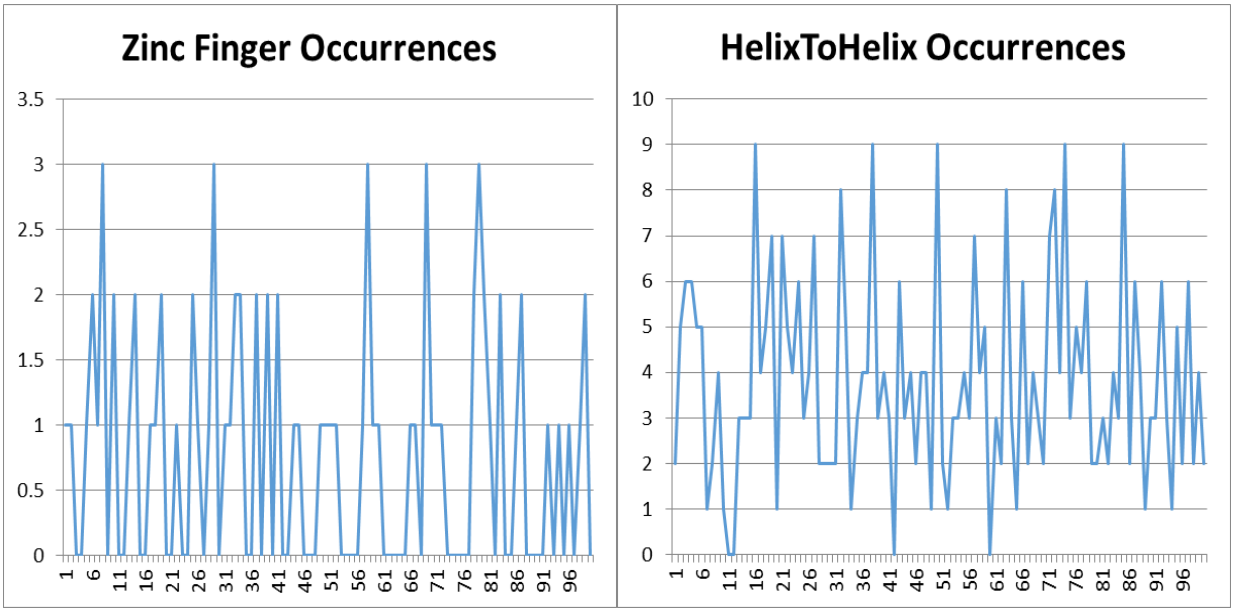

Figure 4.11. Random distribution of the common pattern of the two datasets

The average number of occurrences of group "C24" in new reconstruct graph is 0.77 which is far less than the original 4 times, the standard deviation is 0.885974; the average

number of occurrences of group “C19, C33, C34, C36, C37” in new reconstruct graph is 2.44which is far less than the original 17 times, the standard deviation is 1.370400.

We also download a new dataset containing 40 PDB structures from “Protein Data Bank” which is the X-ray resolution less than 1.5 Å involving both DNA and protein chains. We calculated the number of occurrences of our common group and get result as Figure 4.12:

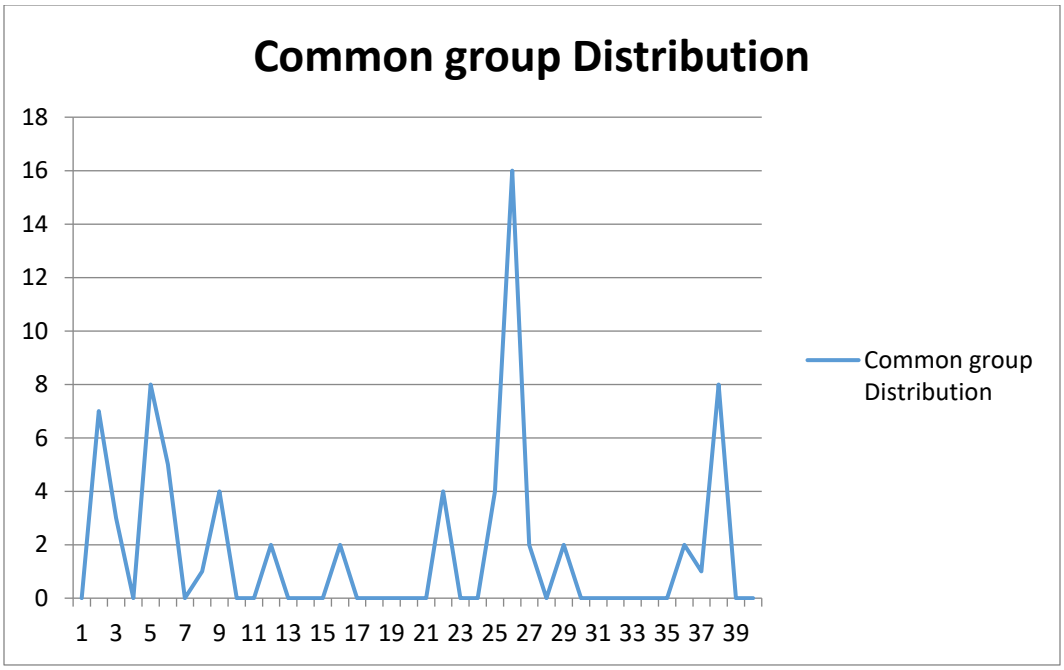

Figure 4.12. Distribution of the common pattern of the two datasets in new 40 PDB complexes

The total number of occurrences is 71. Then we do the same randomly distribution experiment and get the total number of occurrence result is shown in Figure 4.13:

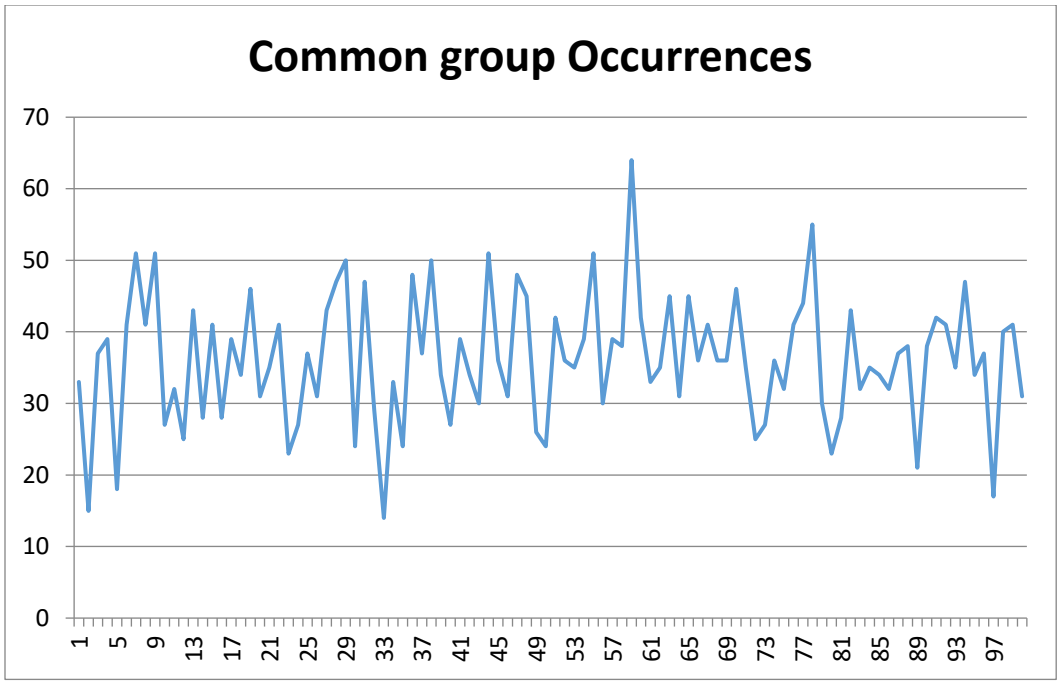

Figure 4.13. Random distribution of the common pattern of the two datasets in new 40 PDB complexes

The average number of occurrences of common group in the randomly reconstruct graph is 36.16 which is far less than the original 71 times, the standard deviation is 8.982272. All these results confirmed the statistical significance of our motifs by calculating the frequencies of occurrences of our motifs in various random distribution situations.

4.5. Apply the Proposed Method to Other Type of Interactions Part II

Protein-protein interactions play important roles in varied biological processes. It has been shown that certain residues at the protein-protein interfaces, contribute more significantly to binding affinity than others. These residues are called hotspots. Generally, a residue is defined as hotspot is its mutation to Alanine results in a decrease of at least 2.0kccal/mol in binding free energy ($\Delta G=Gmut-Gwt$), where Gwt is the binding free energy upon complex formation of wild-type proteins and Gmut is the binding free energy upon complex formation of alanine mutated proteins.

Using biological methods to determine which residues are hot spots can be costly and time consuming. Recent advances in computational approaches to predict hot spots have incorporated a myriad of features and have shown increasing predictive successes. In this part, we apply our new structural method to predict hot spots in protein-protein interaction. We obtained datasets of experimentally varified hotspots, and then used our method to identify crucial common patterns in protein-protein interfaces. Then, compared the patterns with experimentally varified hotspots. We also compared our method with other computational hotspots prediction method to assess the effectiveness of the method.

4.5.1. Dataset Preparation

First, we collected datasets of hotspots from two resources: The Alanine Scanning Energetics Database (ASEdb) which contains results from alanine scanning analysis and the Binding Interface Database (BID), which contains experimentally verified hot spots extracted from literature. The original datasets contained thousands of hotspot records. We filtered out the records that didn't have a PDB id associated with them, and if multiple PDB and PDB ids were associated with one record, we only kept the one with most reliable label of hotspots in its residue chains. At the end, we obtained 19 protein-protein complexes from ASEdb and 29 from BID. Some of these protein-protein interfaces correspond to intermolecular interactions, others intramolecules. Table 4.27 shows the numbers of complexes in each of the categories.

Table 4.27. Number of complexes in each category

	# of intermolecular complexes	**# of intra-molecular complexes**
ASEdb	11	8
BID	22	7

4.5.2. Graph Construction

The protein-protein interfaces in the protein-protein complexes obtained from the previous section were represented using graphs as follows. For each complex, we used NACCESS software to calculate the accessible surface area (ASA) of each amino acid in both bound and unbound states. We collected the interface residues whose ASA in unbound state was at least 1Å more than that in bound state. Then we applied our graph representation method to construct graphs for the protein-protein interfaces such thatevery node of the graph represented an interface residue, and an edge was added between two nodes if the corresponding residues were at a distance less than 10Å. Each node was labeled with its residue type. Each edge was also associated with a type label, where type 0 was edges between nodes from the same protein chain, and type 1 edges between nodes from different protein chains.

4.5.3. Calculation of Common Sub-graphs

Since intermolecular and intramolecular interactions may utilize different structural patterns to facilitate the protein-protein interactions. We searched for common structural motifs for intermolecular and intramolecular interfaces separately. We also did that separately for each of the databases, ASEdb and BID, since they represented hotspots obtained using different methods. For each category of complexes, we discovered common sub-graphs for every pairs of graphs using the maximal clique algorithm. Table 28 showed the numbers of common structural motifsfound in the four categories.

Table 4.28. Numbers of common motif in each category before filtering

	# of common motifs in intermolecular complexes	# of common motifs in intra-molecular complexes
ASEdb	25	135
BID	292	37

Then we filtered the motifs in three steps. First, we removed the common motifs without cross-interface edges, i.e., to remove motifs that had all verticesfrom one protein chain. Second, we removed the common motifs that occur in less than 3 interfaces. Third, we removed duplicated motifs in each category. The final common motif numbers of each category are shown in Table 4.29.

Table 4.29. Final number of common motifs in each category

	# of common motifs in intermolecular complexes	# of common motifs intra-molecular complexes
ASEdb	0	4
BID	23	0

4.5.4. Verification of Our Common Motifs

4.5.4.1. Compare the Discovered Motifs with Hotspots Identified by Other Methods

First, we compare motifs discovered by our methods with the hotspots identified by two computational methods, namely FoldX [7] and Hotsprint [8]. FoldX calculates the free energy of a macromolecule based on its high-resolution 3D structure and uses that toevaluate how a residue's mutation affects the stability, folding and dynamics of the protein structure [7].

Hotsprint is a database of computational hotspots in protein interfaces which use conservation, buried accessible solvent area (ASA) and other important features to predicting the hotspots [8].

Tables 4.30 and Table 4.31 (columns 2, 4, and 5) show that almost all residues on the structural motifs discovered by our method were also predicted to be hotspots by both FoldX and Hotsprint. We also compared the motifs discovered by our method with experimentally verified hotspots in ASEdb and BID. Tables 4.30 and 4.31 (columns 2, and 3) show that some of the residues on our structural motifs were annotated as hotspots in the databases. It is worth noting that many of the residues on our structural motifs were not annotated as hotspots in ASEdb or BID. One possible reason is that the hotspots collected in ASEdb and BID are only a subset of the true hotspots, since only a small fraction of the hotspots have been verified with experiments.

Table 4.30. Result in intra-molecular dataset of ASEdb

PDB ID	# of residues covered by our motifs	# of residues are marked as hotspot in the databases	# of residues are marked as hotspot in method FoldX	# of residues are marked as hotspot in method Hotsprint
1MNM	8	1	8	8
1WAP	3	0	3	3
1BDT	3	0	3	3
1CDC	7	2	6	5

Table 4.31. Result in inter-molecular dataset of BID

PDB ID	# of residues covered by our motifs	# of residues are marked as hotspot in the databases	# of residues are marked as hotspot in method FoldX	# of residues are marked as hotspot in method Hotsprint
1J2X	2	2	2	2
1LQB	6	0	6	6
1F3U	6	0	6	6
1JAT	3	3	3	3
1XDA	5	3	3	3
1CDL	3	2	2	3
1UB4	4	2	4	2
1LEW	1	1	1	1
1NFI	3	1	2	3
1JMA	1	1	0	0

4.5.4.2. Evolutionary Conservation Analysis of the Structural Motifs

Conservation is considered to be an important feature of hotspots in protein-protein interaction. So, we calculated the conservation scores for the residues in the motifs we found. The conservation scores of amino acid residues were calculated using the ConSurf server [8]. For each input protein structure, the server reports normalized conservation scores for all amino acid residues, such that the average is zero and standard deviation is 1. Lower conservation scores correspond to higher conservation levels. Table 4.32 shows the conservation scores for the motifs in the intra-molecular dataset of ASEdb. Each row in the table corresponds to one occurrence of the motif in the protein-protein interfaces. Table 4.33 shows the conservation scores for the motifs on the intermolecular dataset of BID.

Table 4.32. Conservation score in intra-molecular dataset of ASEdb

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1WAP	ILE_B44	-0.335(7)
	LEU_B38	-0.718(9)
	LEU_A24	-0.684(9)
1BDT	PHE_A10	-0.019(5)
	ILE_B37	-0.836(8)
	LEU_B19	-0.286(6)
	VAL_B22	-0.180(6)
	LEU_B12	-0.241(6)
	LEU_A12	-0.241(6)

Table 4.32. Conservation score in intra-molecular dataset of ASEdb (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1MNM	LEU_B59	-0.349(6)
	PHE_A77	-0.086(5)
	LEU_A59	-0.349(6)
	ILE_B80	0.075(5)
	LEU_A89	1.917(1)
	ILE_A43	-0.902(8)
	LEU_A60	-0.210(6)
	LEU_A61	-0.350(6)
	LEU_B89	1.917(1)
	PHE_B77	-0.086(5)
	VAL_A62	-0.322(6)
	LEU_B50	-1.229(9)
	VAL_B62	-0.322(6)
	LEU_B61	-0.350(6)
	ILE_A80	0.075(5)

Table 4.32. Conservation score in intra-molecular dataset of ASEdb (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1CDC	LEU_B68	-1.159(9)
	LEU_A16	-1.248(9)
	ILE_A65	-1.294(9)
	ILE_A97	-0.817(8)
	LEU_B16	-1.248(9)
	VAL_A39	-0.604(7)
	LEU_B38	0.006(5)
	LEU_A38	0.006(5)
	VAL_B39	-0.604(7)
	LEU_A68	-1.159(9)
	LEU_B10	-0.848(8)
	ILE_B65	-1.294(9)
	PHE_B49	2.230(1)
	PHE_A49	2.230(1)
	ILE_A14	-0.807(8)

Table 4.33. Conservation score in inter-molecular dataset of BID

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1J2X	LEU_B953	
	PHE_A513	-0.170(6)
	LEU_B957	
	LEU_A474	-0.308(6)
	LEU_A473	-1.236(9)
	LEU_A497	-0.917(8)
	LEU_A493	-0.225(6)
1LQB	ALA_B107	-1.121(9)
	LEU_C188	-1.185(9)
	LEU_B101	-0.982(8)
	LEU_C178	-1.221(9)
	ILE_C180	-1.192(9)
	LEU_B103	-0.715(7)
	LEU_C158	-1.292(9)
	ALA_B100	-0.788(8)
	LEU_B104	-0.819(8)
	VAL_B73	-0.884(8)
	LEU_C163	-0.488(7)
	VAL_C166	-0.250(6)

Table 4.33. Conservation score in inter-molecular dataset of BID (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1F3U	LEU_A106	0.698(3)
	ILE_B126	-0.823(8)
	LEU_B100	-0.993(9)
	TRP_A31	-1.069(9)
	LEU_B47	-0.791(8)
	ALA_B137	-1.186(9)
	ALA_B45	-0.487(7)
	PHE_B31	-0.912(8)
	PHE_B146	-1.095(9)
	VAL_A95	-1.094(9)
	LEU_A7	-0.370(6)
	VAL_B18	-0.823(8)
	LEU_A9	-0.328(6)
	ILE_B28	-0.533(7)
	LEU_A20	-0.995(9)
	VAL_B140	-0.953(8)
1JAT	LEU_B11	-1.034(9)
	LEU_B30	-0.883(8)
	LEU_A83	0.217(4)
	LEU_B14	-1.035(9)

Table 4.33. Conservation score in inter-molecular dataset of BID (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1XDA	LEU_A13	0.792(1)
	LEU_B11	-1.121(9)
	LEU_A16	-0.640(8)
	VAL_B18	-0.224(6)
	ALA_B14	-0.922(9)
	LEU_B17	1.465(1)
	LEU_B15	-1.016(9)
1CDL	VAL_E807	
	TRP_E800	
	ILE_E810	
	LEU_A39	-0.864(7)
	ALA_A128	-0.666(7)
	VAL_A136	-0.972(8)
1UB4	LEU_C458	-1.238(9)
	LEU_B247	-0.095(5)
	TRP_C473	0.068(5)
	LEU_C455	-1.297(9)
	LEU_A109	-1.077(9)
	LEU_A47	-0.095(5)
	VAL_A78	-0.129(5)
	PHE_A60	-0.750(8)

Table 4.33. Conservation score in inter-molecular dataset of BID (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
	VAL_B278	-0.129(5)
	ALA_A31	-0.760(8)
	ILE_A110	-0.380(6)
	ILE_B310	-0.380(6)
	VAL_A15	-0.259(6)
	VAL_A108	0.205(4)
	VAL_C459	-0.541(7)
1LEW	VAL_A158	-0.501(7)
	VAL_B7	
	LEU_A122	-0.622(7)
	ILE_A116	-0.308(6)
1NFI	ALA_A249	-1.110(9)
	LEU_A215	-1.080(9)
	VAL_B313	-0.550(7)
	ALA_B311	-0.965(9)
	VAL_A251	-0.112(5)
	LEU_B272	-0.929(9)
1JMA	ALA_A7	0.559(3)
	LEU_B49	1.086(2)
	LEU_A4	2.778(1)

These results show that in almost all the residues involved in the motifs have very high conservation scores. This suggests the structural and functional importance of the common motifs discovered by our method.

4.5.4.3. Statistical Significance of the Discovered Motifs

In this project, we have discovered 4 structural motifs in intramolecular protein-protein interfaces and 21 in intermolecular protein-protein interfaces. In this section, we will assess the statistical significances of these motifs. In other words, we assess whether the occurrences of these motifs in the protein-protein interfaces is the result of a random process or the consequence of the fact that they are crucial for protein-protein interactions. To answer this question, we used two larger datasets generated in [9], which contained 174 and 429 protein-protein complexes respectively. We separated the complexes in the datasets into intra-molecular and intermolecular interfaces. First, we tallied how many times each of the motifs was observed in these databases. Then we used a bootstrapping method to estimate how many times these motifs would occur in the interfaces due to random process. The results are shown in Tables 4.34 - 4.37. For each motif, we performed the bootstrapping experiment 100 times, and used t-test to test the difference between the time of occurrences observed in the real interfaces (column 3 of the tables) and that of random process (column 4 of the tables). The results showed that for all the motifs, the number of occurrences observed in the real interfaces is much higher than that in random process. Almost all the t-tests have a p value lower than 0.0001. This result suggests that the motifs discovered by our method are significantly associated with protein-protein interactions, rather than a result of random process.

Table 4.34. Occurrence and Bootstrapping of 174 datasets (intra part)

Structural Motif ID	Amino Acid Categories of the Vertices	Observed frequencies in the real Protein-Protein interfaces	Average frequencies in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	LEU, LEU, ILE	5	1.14	1.114	Less than 0.0001
2	LEU, LEU, ILE	5	1.73	1.522	Less than 0.0001
3	LEU, LEU, ILE	7	1.37	1.369	Less than 0.0001
4	PHE, LEU, VAL	3	1.17	1.07	Less than 0.0001

Table 4.35. Occurrence and Bootstrapping of 174 datasets (inter part)

Structural Motif ID	Amino Acid Categories of the Vertices	Observed frequencies in the real Protein-Protein interfaces	Average frequencies in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	LEU, LEU, LEU	2	1.28	1.225398	Less than 0.0001
2	LEU, LEU, LEU	0	1.07	1.193776	Less than 0.0001
3	LEU, LEU, LEU	3	1.2	1.104536	Less than 0.0001
4	LEU, LEU, LEU	2	1.29	1.194111	Less than 0.0001
5	LEU, LEU, LEU	2	1.07	1.012472	Less than 0.0001
6	LEU, LEU, LEU	4	0.88	1.041921	Less than 0.0001
7	LEU, LEU, LEU	4	1.19	1.238507	Less than 0.0001
8	LEU, ALA, LEU	1	1	0.979796	1
9	LEU, LEU, VAL	3	0.63	0.820427	Less than 0.0001
10	LEU, LEU, VAL	2	0.68	0.76	Less than 0.0001
11	LEU, LEU, LEU	4	0.92	1.270276	Less than 0.0001
12	LEU, LEU, ALA	1	0.76	0.861626	Less than 0.0001

Table 4.35. Occurrence and Bootstrapping of 174 datasets (inter part) (continued)

Structural Motif ID	Amino Acid Categories of the Vertices	Observed frequencies in the real Protein-Protein interfaces	Average frequencies in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
13	LEU, LEU, LEU	1	1.06	1.129779	0.5963
14	ALA, TRP, VAL	1	0.19	0.440341	Less than 0.0001
15	ILE, VAL, VAL	0	0.34	0.586856	Less than 0.0001
16	LEU, LEU, PHE	0	0.57	0.751731	Less than 0.0001
17	LEU, LEU, LEU	2	1.17	1.131857	Less than 0.0001
18	LEU, LEU, VAL	2	0.81	0.879716	Less than 0.0001
19	LEU, LEU, VAL	4	0.79	1.061084	Less than 0.0001
20	ALA, VAL, LEU	2	0.49	0.754917	Less than 0.0001
21	LEU, LEU, ALA	2	1.13	1.308854	Less than 0.0001
22	LEU, VAL, ILE	1	0.49	0.714073	Less than 0.0001
23	VAL, LEU, VAL	1	0.77	0.834925	0.0070

Table 4.36. Occurrence and Bootstrapping of 429 datasets (intra part)

Structural Motif ID	Amino Acid Categories of the Vertices	Observed frequencies in the real Protein-Protein interfaces	Average frequencies in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	LEU, LEU, ILE	4	1.72	1.364405	Less than 0.0001
2	LEU, LEU, ILE	20	2.92	1.753169	Less than 0.0001
3	LEU, LEU, ILE	6	2.85	1.669581	Less than 0.0001
4	PHE, LEU, VAL	3	2.12	1.394848	Less than 0.0001

Table 4.37. Occurrence and Bootstrapping of 429 datasets (inter part)

Structural Motif ID	Amino Acid Categories of the Vertices	Observed frequencies in the real Protein-Protein interfaces	Average frequencies in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	LEU, LEU, LEU	6	1.39	1.475771	Less than 0.0001
2	LEU, LEU, LEU	1	1.14	1.039423	0.1809
3	LEU, LEU, LEU	5	1.04	1.18254	Less than 0.0001
4	LEU, LEU, LEU	4	0.98	1.113373	Less than 0.0001
5	LEU, LEU, LEU	4	0.98	0.927146	Less than 0.0001
6	LEU, LEU, LEU	2	1.27	1.317991	Less than 0.0001
7	LEU, LEU, LEU	4	1.23	1.231706	Less than 0.0001
8	LEU, ALA, LEU	0	0.83	0.990505	Less than 0.0001
9	LEU, LEU, VAL	1	0.68	0.881816	0.0004
10	LEU, LEU, VAL	1	0.71	0.851998	0.0009
11	LEU, LEU, LEU	1	0.94	0.925419	0.5181

Table 4.37. Occurrence and Bootstrapping of 429 datasets (inter part) (continued)

Structural Motif ID	Amino Acid Categories of the Vertices	Observed frequencies in the real Protein-Protein interfaces	Average frequencies in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
12	LEU, LEU, ALA	2	0.72	0.800999	Less than 0.0001
13	LEU, LEU, LEU	2	0.88	0.992774	Less than 0.0001
14	ALA, TRP, VAL	2	0.14	0.424735	Less than 0.0001
15	ILE, VAL, VAL	1	0.35	0.554527	Less than 0.0001
16	LEU, LEU, PHE	5	0.5	0.754983	Less than 0.0001
17	LEU, LEU, LEU	2	0.71	0.930537	Less than 0.0001
18	LEU, LEU, VAL	2	0.77	0.914932	Less than 0.0001
19	LEU, LEU, VAL	0	1.11	1.085311	Less than 0.0001
20	ALA, VAL, LEU	1	0.33	0.617333	Less than 0.0001
21	LEU, LEU, ALA	4	0.81	0.783518	Less than 0.0001
22	LEU, VAL, ILE	0	0.6	0.812404	Less than 0.0001
23	VAL, LEU, VAL	1	0.87	1.035905	0.2121

For each occurrence of the motif in these two datasets, we also analyze the conservation of the residues corresponding to the vertices of the motifs. The results are shown in Tables 4.38 - 4.41. The results show that these residues have very high levels of conservations, indicating the functional importance of the motifs.

Table 4.38. Conservation scores of residues in 174 datasets (intra part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1MI3	ILE-B320	1.694(1)
	LEU-B175	-0.755(7)
	LEU-B177	0.104(5)
	LEU-A177	0.104(5)
	ILE-A320	1.694(1)
	LEU-A175	-0.755(7)
1W0I	LEU-A172	1.171(1)
	LEU-B243	1.317(1)
	ILE-B167	0.799(3)
	ILE-A167	0.799(3)
	LEU-B172	1.171(1)
	LEU-A243	1.317(1)
1B99	LEU-C39	-0.419(6)
	ILE-C29	-0.084(5)
	LEU-F42	0.402(4)

Table 4.38. Conservation scores of residues in 174 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1BRR	LEU-C95	-0.059(5)
	LEU-A109	1.205(1)
	ILE-C45	-0.319(6)
1HWU	LEU-A13	-0.115(5)
	VAL-A59	0.866(2)
	PHE-B55	0.002(5)
2COG	VAL-A92	-0.968(8)
	LEU-A173	-0.871(8)
	PHE-B53	-0.745(7)
2D4V	ILE-A198	-0.246(6)
	LEU-B179	-0.573(7)
	ILE-B175	0.573(3)
	ILE-A175	0.573(3)
	ILE-B198	-0.246(6)
	LEU-A179	-0.573(7)
1VGQ	PHE-B63	-0.834(9)
	LEU-A167	-0.881(9)
	VAL-A16	-0.965(9)

Table 4.39. Conservation scores of residues in 174 datasets (inter part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
2O8A	LEU-A55	0.042(5)
	LEU-A53	-0.195(6)
	LEU-I15	-1.448(9)
2BE6	LEU-A112	-0.075(5)
	VAL-D1615	-0.577(8)
	VAL-A55	-0.115(5)
1GXS	VAL-D351	-0.179(6)
	LEU-D349	-0.166(6)
	LEU-D344	-0.588(7)
	LEU-C184	0.153(4)
2OCC	VAL-A299	0.368(4)
	ILE-B42	-0.262(6)
	VAL-B38	-1.037(9)
	LEU-B84	-0.896(8)
	LEU-B37	1.328(1)
	ALA-A303	-0.362(6)
	LEU-A324	-0.011(5)
2BL0	ALA-A791	0.662(3)
	LEU-B117	-1.046(8)
	VAL-A795	0.691(3)

Table 4.39. Conservation scores of residues in 174 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
3PCD	TRP-M400	-1.179(9)
	VAL-A17	-0.579(7)
	ALA-A13	-0.964(8)
2PRG	LEU-C633	-0.145(5)
	LEU-C636	-0.933(8)
	LEU-A468	-1.084(9)
1PST	LEU-H12	-1.078(9)
	LEU-M286	-0.526(8)
	VAL-M290	-0.547(8)
	LEU-M275	-0.597(8)
	ALA-H16	-0.220(6)
	LEU-M278	0.164(4)
	LEU-H27	-1.180(9)
	ALA-H13	-1.176(9)
2P1L	LEU-A108	-0.296(6)
	LEU-A130	-1.017(9)
	LEU-B116	1.403(1)
	LEU-A194	0.394(4)
	VAL-A141	-1.007(9)

Table 4.39. Conservation scores of residues in 174 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1NHG	LEU-D399	2.113(1)
	VAL-B213	-0.572(8)
	LEU-B119	-0.755(9)
1DM5	LEU-D101	1.135(2)
	LEU-B80	0.864(2)
	ALA-B77	1.278(1)
1QGE	LEU-D17	-0.558(7)
	LEU-E265	-0.474(7)
	LEU-E314	-1.132(9)
	ALA-D105	-1.196(9)
	LEU-D205	0.004(5)
	LEU-D149	0.240(4)
	LEU-E247	0.185(4)
	LEU-D164	-0.132(5)
	LEU-E292	1.416(1)
	LEU-E286	-0.192(6)

Table 4.40. Conservation scores of residues in 429 datasets (intra part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1G85	VAL-B100	-0.420(6)
	PHE-A132	-1.431(9)
	LEU-B118	0.068(5)
1CWQ	LEU-A92	-1.499(9)
	LEU-B99	-0.630(7)
	ILE-B45	-0.191(6)
	ILE-A45	-0.191(6)
2FYI	ILE-A112	-0.591(7)
	LEU-A126	-0.564(7)
	ILE-B112	-0.591(7)
	LEU-B126	-0.564(7)
	LEU-B124	-0.664(8)
	LEU-A236	-0.454(7)
	LEU-B236	-0.454(7)
1I0Z	ILE-A37	0.085(5)
	LEU-A65	-0.833(8)
	LEU-B254	0.099(5)
	ILE-B37	0.085(5)
	LEU-B65	-0.833(8)
	LEU-A254	0.099(5)

Table 4.40. Conservation scores of residues in 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
2NUU	VAL-C50	0.653(3)
	LEU-C54	0.248(4)
	PHE-B247	1.706(1)
1AUW	ILE-D225	0.787(2)
	LEU-C440	1.479(1)
	LEU-D227	1.676(1)
1TW2	ILE-B126	-0.680(7)
	ILE-A74	0.650(3)
	LEU-B67	-0.625(7)
	LEU-A301	-1.183(8)
	LEU-A304	-0.925(8)
	LEU-A76	-0.315(6)
	ILE-A126	-0.680(7)
	LEU-B304	-0.925(8)
	LEU-B22	0.157(5)
	ILE-B74	0.650(3)
	LEU-A67	-0.625(7)
	LEU-B76	-0.315(6)
	LEU-B301	-1.183(8)
	LEU-B122	-1.180(8)
	LEU-A22	0.157(5)

Table 4.40. Conservation scores of residues in 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1O5O	LEU-B11	-0.889(8)
	ILE-A67	0.215(4)
	LEU-A48	0.063(5)
2EGH	ILE-A255	-0.487(7)
	LEU-A270	-0.405(6)
	LEU-B270	-0.405(6)
1V2I	VAL-A187	0.960(1)
	LEU-A555	-0.997(9)
	PHE-B563	-0.440(7)
1EWK	ILE-A120	-0.783(8)
	ILE-B120	-0.783(8)
	LEU-A174	-0.842(8)
	LEU-B174	-0.842(8)
1P0K	ILE-A74	-0.843(8)
	ILE-B74	-0.843(8)
	LEU-A70	1.095(1)
	LEU-B70	1.095(1)
1HKV	LEU-A347	-1.194(9)
	LEU-B347	-1.194(9)
	ILE-B343	-0.543(7)

Table 4.40. Conservation scores of residues in 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1AB8	LEU-B877	-0.696(8)
	LEU-A915	-0.899(8)
	ILE-B1010	-1.055(9)
1XI9	ILE-A271	-0.614(7)
	LEU-A268	-0.509(7)
	LEU-B133	0.499(3)

Table 4.41. Conservation scores of residues in 429 datasets (inter part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1M56	LEU-B123	-0.998(8)
	LEU-B120	-0.884(8)
	LEU-A342	0.773(2)
1PYT	LEU-B280	1.099(2)
	LEU-B125	0.189(4)
	LEU-A86	-0.382(6)
1AZS	LEU-B912	-0.828(8)
	LEU-A497	-0.758(8)
	LEU-B915	-0.995(9)
1NF3	LEU-B67	-1.056(8)
	LEU-D208	-0.750(9)
	LEU-B70	-0.955(8)

Table 4.41. Conservation scores of residues in 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1MF8	LEU-A369	-0.849(8)
	VAL-A368	-0.181(6)
	LEU-B29	-0.828(7)
1Z7M	LEU-E84	-0.308(6)
	PHE-E188	0.310(4)
	LEU-A192	0.501(3)
1JWI	LEU-B79	-0.389(6)
	LEU-B92	-0.539(7)
	LEU-A70	-1.101(8)
	PHE-A101	-0.389(6)
	LEU-B87	0.562(3)
1AIG	ALA-P13	-1.152(9)
	LEU-O286	-0.647(9)
	LEU-P27	-1.166(9)
	LEU-O275	-0.579(8)
	LEU-P24	-1.121(9)
	LEU-P12	-0.968(8)
1C0T	LEU-B289	-0.680(9)
	VAL-A496	-0.501(8)
	ALA-A534	1.042(1)

Table 4.41. Conservation scores of residues in 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1JFF	VAL-B260	-0.735(8)
	TRP-A407	-0.859(8)
	ALA-B256	-0.596(8)
	VAL-B257	-0.733(8)
1BML	LEU-C292	1.217(1)
	LEU-C314	0.587(3)
	LEU-A626	1.484(1)
1OVL	VAL-B369	-0.262(6)
	VAL-A373	0.810(2)
	ILE-A500	0.238(4)
1HXM	LEU-A141	-0.859(9)
	ALA-A139	-0.954(9)
	LEU-B153	-0.758(8)
2BTW	PHE-B36	-0.444(6)
	LEU-A45	-0.630(7)
	LEU-B46	3.125(1)
	LEU-B45	-0.692(7)
	LEU-A46	3.154(1)
	PHE-A36	-0.470(6)

Table 4.41. Conservation scores of residues in 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	Conservation score of vertexes (9 - conserved, 1 - variable)
1NBU	VAL-B18	-1.058(9)
	VAL-C55	-0.070(5)
	LEU-C48	-0.554(7)
1AR1	LEU-B100	-1.015(8)
	LEU-B97	-0.651(7)
	LEU-A334	0.253(4)
	ALA-A338	-0.055(5)
	LEU-A342	0.058(5)
1UKV	LEU-G233	-0.449(7)
	LEU-Y193	1.797(1)
	VAL-Y191	0.149(4)
1SB2	VAL-B85	-0.309(6)
	LEU-A70	-1.039(9)
	LEU-B77	-0.117(5)
	LEU-B90	-0.536(7)
1XDK	LEU-B371	-0.844(9)
	LEU-A425	-1.296(9)
	LEU-A424	-1.191(9)
	ALA-A421	-1.046(8)
	LEU-B349	-0.695(8)
	PHE-A420	-1.278(9)

For the further verification, we used FoldX and Hotsprint to predict hotspots on the two larger datasets. The results Tables 4.42 - 4.45 show that the majority of the residues on the motifs were predicted to be hotspots by both methods.

Table 4.42. Foldx and Hotsprint results of 174 datasets (intra part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1MI3	ILE-B320	2.92288	H
	LEU-B175	2.54529	H
	LEU-B177	1.6709	H
	LEU-A177	1.60163	H
	ILE-A320	2.80066	H
	LEU-A175	2.1897	H
1W0I	LEU-A172	2.662	H
	LEU-B243	3.85743	H
	ILE-B167	3.17779	H
	ILE-A167	2.97933	H
	LEU-B172	2.21563	H
	LEU-A243	3.98771	H

Table 4.42. Foldx and Hotsprint results of 174 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1B99	LEU-C39	4.09287	H
	ILE-C29	3.5827	H
	LEU-F42	4.49903	H
1BRR	LEU-C95	1.60257	NH
	LEU-A109	1.06434	NH
	ILE-C45	1.97384	H
1HWU	LEU-A13	0.89307	H
	VAL-A59	1.94013	H
	PHE-B55	2.12472	NH
2COG	VAL-A92	2.07821	H
	LEU-A173	2.10625	H
	PHE-B53	2.76446	H

Table 4.42. Foldx and Hotsprint results of 174 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
2D4V	ILE-A198	4.64675	H
	LEU-B179	4.15857	H
	ILE-B175	3.72897	H
	ILE-A175	3.9328	H
	ILE-B198	4.65347	H
	LEU-A179	3.94863	H
1VGQ	PHE-B63	2.90974	H
	LEU-A167	1.72093	H
	VAL-A16	1.39992	H

Table 4.43. Foldx and Hotsprint results of 174 datasets (inter part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
2O8A	LEU-A55	4.75152	H
	LEU-A53	3.22556	H
	LEU-I15	2.05132	H

Table 4.43. Foldx and Hotsprint results of 174 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
2BE6	LEU-A112	3.19144	H
	VAL-D1615	0.55199	NH
	VAL-A55	2.03257	H
1GXS	VAL-D351	3.29978	H
	LEU-D349	2.43339	H
	LEU-D344	3.32737	H
	LEU-C184	3.25701	H
2OCC	VAL-A299	0.899222	NH
	ILE-B42	1.10496	H
	VAL-B38	1.3177	H
	LEU-B84	2.5241	H
	LEU-B37	0.788803	NH
	ALA-A303	0	H
	LEU-A324	3.39641	H

Table 4.43. Foldx and Hotsprint results of 174 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
2BL0	ALA-A791	0	H
	LEU-B117	3.18326	H
	VAL-A795	1.42079	H
3PCD	TRP-M400	3.28324	H
	VAL-A17	1.65332	H
	ALA-A13	0	H
2PRG	LEU-C633	2.88897	H
	LEU-C636	1.78349	NH
	LEU-A468	3.17128	H
2P1L	LEU-A108	1.59735	H
	LEU-A130	2.47983	H
	LEU-B116	3.51503	H
	LEU-A194	0.666135	NH
	VAL-A141	1.95075	H

Table 4.43. Foldx and Hotsprint results of 174 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1PST	LEU-H12	-0.333749	NH
	LEU-M286	1.92521	H
	VAL-M290	0.894359	NH
	LEU-M275	2.76555	H
	ALA-H16	0	NH
	LEU-M278	1.47245	NH
	LEU-H27	2.8672	H
	ALA-H13	0	H
1NHG	LEU-D399	4.02936	H
	VAL-B213	3.30741	H
	LEU-B119	2.7882	H
1DM5	LEU-D101	2.8537	H
	LEU-B80	2.3943	H
	ALA-B77	0	NH

Table 4.43. Foldx and Hotsprint results of 174 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1QGE	LEU-D17	1.68968	H
	LEU-E265	3.68743	H
	LEU-E314	2.76683	H
	ALA-D105	0	NH
	LEU-D205	1.63874	H
	LEU-D149	2.86152	H
	LEU-E247	2.48624	H
	LEU-D164	3.02598	H
	LEU-E292	1.16693	H
	LEU-E286	2.75518	H

Table 4.44. Foldx and Hotsprint results of 429 datasets (intra part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1G85	VAL-B100	3.32671	H
	PHE-A132	5.4911	H
	LEU-B118	3.84064	H
2FYI	ILE-A112	2.25389	H
	LEU-A126	2.81702	H
	ILE-B112	3.02598	H
	LEU-B126	3.26232	H
	LEU-B124	3.54604	H
	LEU-A236	1.77664	NH
	LEU-B236	1.51826	H
1CWQ	LEU-A92	-69.8588	H
	LEU-B99	-38.1616	H
	ILE-B45	-59.2146	H
	ILE-A45	-46.5357	H

Table 4.44. Foldx and Hotsprint results of 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1I0Z	ILE-A37	2.24517	H
	LEU-A65	3.5992	H
	LEU-B254	1.24123	H
	ILE-B37	2.06719	H
	LEU-B65	3.23862	H
	LEU-A254	1.23857	H
2NUU	VAL-C50	1.88295	H
	LEU-C54	2.4173	H
	PHE-B247	1.06479	NH
1AUW	ILE-D225	1.72257	H
	LEU-C440	0.939303	NH
	LEU-D227	0.825853	NH
1O5O	LEU-B11	2.85004	H
	ILE-A67	3.01077	H
	LEU-A48	2.56391	H

Table 4.44. Foldx and Hotsprint results of 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1TW2	ILE-B126	2.87936	H
	ILE-A74	2.53761	H
	LEU-B67	3.696	H
	LEU-A301	2.90852	H
	LEU-A304	1.81494	H
	LEU-A76	3.42717	H
	ILE-A126	3.18988	H
	LEU-B304	2.04903	H
	LEU-B22	1.41521	NH
	ILE-B74	2.92941	H
	LEU-A67	3.12384	H
	LEU-B76	3.69269	H
	LEU-B301	2.95945	H
	LEU-B122	2.01551	H
	LEU-A22	2.32137	H

Table 4.44. Foldx and Hotsprint results of 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
2EGH	ILE-A255	3.05018	H
	LEU-A270	1.84109	H
	LEU-B270	3.41902	H
1V2I	VAL-A187	-0.207707	H
	LEU-A555	2.78884	H
	PHE-B563	4.86801	H
1EWK	ILE-A120	2.11047	H
	ILE-B120	2.21801	H
	LEU-A174	3.16704	H
	LEU-B174	3.15242	H
1P0K	ILE-A74	1.75001	H
	ILE-B74	1.81817	H
	LEU-A70	2.8364	H
	LEU-B70	2.8744	H

Table 4.44. Foldx and Hotsprint results of 429 datasets (intra part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1HKV	LEU-A347	3.21508	H
	LEU-B347	3.67745	H
	ILE-B343	2.42404	H
1AB8	LEU-B877	2.28387	H
	LEU-A915	2.92425	H
	ILE-B1010	2.5096	H
1XI9	ILE-A271	2.06895	H
	LEU-A268	3.1054	H
	LEU-B133	2.53108	H

Table 4.45. Foldx and Hotsprint results of 429 datasets (inter part)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1M56	LEU-B123	1.38927	H
	LEU-B120	1.19847	NH
	LEU-A342	1.16217	NH

Table 4.45. Foldx and Hotsprint results of 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1PYT	LEU-B280	2.58962	H
	LEU-B125	3.50036	H
	LEU-A86	1.53339	H
1AZS	LEU-B912	2.68622	H
	LEU-A497	3.1188	H
	LEU-B915	2.49169	H
1NF3	LEU-B67	1.90269	H
	LEU-D208	3.59395	H
	LEU-B70	0.496832	NH
1MF8	LEU-A369	1.61408	H
	VAL-A368	2.00366	H
	LEU-B29	3.68364	H
1Z7M	LEU-E84	2.15318	H
	PHE-E188	2.27782	H
	LEU-A192	1.55385	H

Table 4.45. Foldx and Hotsprint results of 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1JWI	LEU-B79	3.83968	H
	LEU-B92	3.22457	H
	LEU-A70	4.1262	H
	PHE-A101	5.32497	H
	LEU-B87	1.65859	H
1AIG	ALA-P13	0	H
	LEU-O286	2.41348	H
	LEU-P27	3.43386	H
	LEU-O275	3.67513	H
	LEU-P24	2.15223	H
	LEU-P12	-0.25158	NH
1C0T	LEU-B289	2.77368	H
	VAL-A496	2.50007	H
	ALA-A534	0	NH

Table 4.45. Foldx and Hotsprint results of 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1JFF	VAL-B260	-3.16249	H
	TRP-A407	-1.52793	NH
	ALA-B256	0	H
	VAL-B257	-1.47161	NH
1BML	LEU-C292	-1.94925	H
	LEU-C314	2.31688	H
	LEU-A626	-0.75152	H
1OVL	VAL-B369	2.23022	H
	VAL-A373	1.38848	H
	ILE-A500	1.66276	H
1HXM	LEU-A141	2.70031	H
	ALA-A139	0	H
	LEU-B153	2.91486	H

Table 4.45. Foldx and Hotsprint results of 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
2BTW	PHE-B36	3.10236	H
	LEU-A45	2.88869	H
	LEU-B46	1.1247	NH
	LEU-B45	2.88703	H
	LEU-A46	1.06047	NH
	PHE-A36	3.26134	H
1NBU	VAL-B18	-0.883619	NH
	VAL-C55	0.564909	NH
	LEU-C48	0.149154	NH
1AR1	LEU-B100	2.29058	H
	LEU-B97	2.1262	NH
	LEU-A334	2.32249	H
	ALA-A338	0	H
	LEU-A342	0.979486	NH

Table 4.45. Foldx and Hotsprint results of 429 datasets (inter part) (continued)

PDB ID	Identity and PDB index of vertex (which is an amino acid)	FoldX score of vertexes (DDG upon alanine mutation in Kcal/mol)	HotSprint result of vertex (H for Hotspot and NH for not Hotspot)
1UKV	LEU-G233	4.4628	H
	LEU-Y193	2.5709	H
	VAL-Y191	1.87093	NH
1SB2	VAL-B85	1.69258	NH
	LEU-A70	3.55141	H
	LEU-B77	3.3493	H
	LEU-B90	1.22833	H
1XDK	LEU-B371	3.11804	H
	LEU-A425	1.90253	H
	LEU-A424	2.9613	H
	ALA-A421	0	H
	LEU-B349	1.68549	H
	PHE-A420	2.48183	H

4.6. Apply the Proposed Method to Other Type of Interactions Part III

4.6.1. Dataset Preparation

In this part our study focused on the interaction between protein and a significant ligand, ATP. ATP is a complex organic chemical nucleotide which plays an important role in many biological processes as a coenzyme interacting with proteins [8]. ATP is also considered to be

the energy currency of life. We collected all protein-ATP complexes from PDB (https://www.rcsb.org/pdb) that were obtained by X-ray method with a resolution less than 1.5 Å. Then we used the PISCES web server [13] to filter the complexes so that pair-wise identities between protein chains were less than 25%. We also removed the entries with the length of the protein chain less than 40. Finally, the protein-ligand dataset included 18 protein-ligand interfaces in 17 PDB complexes (Table 4.46)

Table 4.46. Protein-ligand domains

PDB ID	Chain ID	PDB ID	Chain ID
1OBD	A	4ZQX	A
1XDN	A	5CU6	C
2C01	X	5ETN	A
3GAH	A	5GQI	A
3QXC	A	5HNV	A
4AFF	A	5J1S	A
4B1Y	B	5LVO	A
4C5C	A	5XD4	A
4NDO	A	5XD4	A

4.6.2. Graph Construction

First, we divided ATP into three ATP subgroups: phosphoric acid (AT1), sugar (AT2) and base (AT3). These subgroups were considered building blocks of the ATP, similar to the amino acids of proteins. For each protein-ATP complexes, we extracted the amino acids and ATP subgoups that were located at the protein-ATP interfaces. An amino acid residue was considered at the protein-Ligand interfaces if the closest distance between its side chain and ATP was less than 5 Å. An ATP subgroup was considered at the interface if the closet distance

between it and any amino acid side chain was less than 5 Å. The threshold of 5 Å was chosen because interactions with distances longer than that wouldn't have significant contributions to the binding affinity. Then, each protein-Ligand interface was represented as a graph, with each residue (or ATP subgroup) being represented as a vertex and an edge being added between two vertices if the distance between them was less than 5 Å. The edges have two types 1 and 0, with 1 representing edges crossing the interface (i.e., they represent contacts between amino acids and ATP）, and 0 being edegs that did not cross the interfaces. Thus, we obtained 18 graphs corresponding to the protein-ATP interfaces. Each vertex was labeled with the category of its amino acid (or ATP subpart), and each edge was labeled with its length and type.

4.6.3. Calculation of Common Sub-graphs

A 3-dimenstional structural motif across the protein-Ligand interface can be defined by the set of amino acids and ATP subgroups involved in the motif and all pairwise distances between them. Such a motif can be represented as a clique, a graph in which every vertex is adjacent to every other vertex. Each vertex of the clique is labeled with its amino acid category (or ATP subpart type) and each edge is labeled with the distance and type between its vertices. Therefore, the problem of discovering common structural motifs at the protein-Ligand interfaces can be transformed into finding common cliques. In this work, we are interested in finding maximal cliques that are not a sub-graph of other cliques. We used the product graph method [14] to find maximal cliques shared by a pair of graphs. First, for each pair of graph G1 and G2, a product graph, G1XG2, was built as follows: (1) For every vertex, named v1, of G1 and every vertex, named u1, of G2, if the labels of v1 and u1 are the same, then a vertex, named (v1, u1), was created for the product graph G1XG2; (2) For every pair of vertices of G1XG2, named (v1, u1) and (v2, u2), an edge was created between them if (a) there was an edge between v1 and v2

in G1, (b) there was an edge between u1 and u2 in G2, and (c) $|edge_{v1-v2} - edge_{u1-u2}| \leq 1\text{Å}$, where $edge_{v1-v2}$ is the distance between v1 and v2. Then, each maximal clique in the product graph G1XG2 corresponded to a common structural motif shared by G1 and G2. For our dataset, we used the product graph method to discover maximal cliques for each pair of graphs. Since the goal was find frequent structural motifs spanning the interfaces, we only kept the maximal cliques that included both amino acids and ATP subparts with the requirement that they must occur in at least three of the protein-Ligand interfaces.

In the protein-ligand dataset, we found 24 common structural motifs that each occurred in at least6 protein-ATP interfaces. These motifs occur a total of 213 times in the 18 protein-ATP interfaces. Each of the motifs contained at least 1 amino acid residues and 1 ATP subgroup. Table 4.47 shows the composition and frequencies of the motifs.

Table 4.47. Structural motifs found in the ATP dataset

Structural Motif ID	Vertex Types	Frequencies	Distribution (# of coverage PDB)
1	ARG, AT1, AT2	9	8
2	ARG, AT1, GLU	6	6
3	ARG, AT1, GLY	10	6
4	ASP, AT1, AT2	7	6
5	AT1, AT2, GLU	8	8
6	AT1, AT2, GLY	14	9
7	AT1, AT2, LYS	10	8
	AT1, AT2, LYS	11	9

Table 4.47. Structural motifs found in the ATP dataset (continued)

Structural Motif ID	Vertex Types	Frequencies	Distribution (# of coverage PDB)
8	AT1, GLY, ILE	7	6
9	AT1, GLY, LYS	9	6
	AT1, GLY, LYS	9	7
	AT1, GLY, LYS	12	6
10	AT1, GLY, VAL	11	7
11	AT1, ILE, LYS	8	7
12	AT1, LEU, LYS	9	7
13	AT2, AT3, ILE	7	6
14	AT2, AT3, LEU	10	8
	AT2, AT3, LEU	6	6
15	AT2, AT3, VAL	7	7
16	AT2, GLY, LYS	12	7
17	AT3, ILE, VAL	8	6
18	AT3, LEU, LEU	8	6
19	AT3, LEU, TYR	6	6
20	AT3, LEU, VAL	9	7

GTP, CTP, and TTP have a structure similar to ATP. They have identical phosphoric acid and sugar subgroups and only differ in the base. We continued to explore whether the motifs discovered above are unique to the protein-ATP interactions or they are common motifs shared in the interactions of GTP, CTP, and TTP with proteins.

We extracted the protein-CTP, protein-GTP, protein-TTP, complexes using the same method as ATP dataset. The numbers of these structures were much less than the protein-ATP

complexes. Thus, in order to obtain sufficient data, we kept all X-ray structures with resolution less than 3 Å. Then we represented the protein-ligand interfaces in these datasets as graphs and counted the occurrence frequencies of the motifs in these protein interfaces. The results are shown in Table 4.48.

Table 4.48. Total occurrence of structural motifs in three new datasets

Dataset	**# of PDB complexes**	**# of motif occurrence**
protein-CTP	37	156
protein-GTP	71	358
protein-TTP	21	104

Tables 4.49 - 4.51 show the composition and frequencies of the motifs in these three datasets.

Table 4.49. Structural motifs found in the CTP dataset

Structural Motif ID	**Vertex Types**	**Frequencies**	**Distribution (# of coverage PDB)**
1	ARG, AT1, AT2	7	7
2	ARG, AT1, GLU	7	5
3	ARG, AT1, GLY	7	6
4	ASP, AT1, AT2	12	12
5	AT1, AT2, GLU	4	4
6	AT1, AT2, GLY	16	12
7	AT1, AT2, LYS	14	13
	AT1, AT2, LYS	17	16

Table 4.49. Structural motifs found in the CTP dataset (continued)

Structural Motif ID	Vertex Types	Frequencies	Distribution (# of coverage PDB)
8	AT1, GLY, ILE	3	3
9	AT1, GLY, LYS	2	2
	AT1, GLY, LYS	5	5
	AT1, GLY, LYS	7	5
10	AT1, GLY, VAL	11	6
11	AT1, ILE, LYS	9	8
12	AT1, LEU, LYS	5	5
13	AT2, AT3, ILE	8	6
14	AT2, AT3, LEU	6	6
	AT2, AT3, LEU	2	2
15	AT2, AT3, VAL	6	4
16	AT2, GLY, LYS	4	3
17	AT3, ILE, VAL	1	1
18	AT3, LEU, LEU	2	1
19	AT3, LEU, TYR	1	1
20	AT3, LEU, VAL	0	0

Table 4.50. Structural motifs found in the GTP dataset

Structural Motif ID	Vertex Types	Frequencies	Distribution (# of coverage PDB)
1	ARG, AT1, AT2	20	15
2	ARG, AT1, GLU	10	7
3	ARG, AT1, GLY	7	7

Table 4.50. Structural motifs found in the GTP dataset (continued)

Structural Motif ID	Vertex Types	Frequencies	Distribution (# of coverage PDB)
4	ASP, AT1, AT2	10	9
5	AT1, AT2, GLU	7	7
6	AT1, AT2, GLY	44	31
7	AT1, AT2, LYS	31	28
	AT1, AT2, LYS	34	33
8	AT1, GLY, ILE	16	12
9	AT1, GLY, LYS	8	8
	AT1, GLY, LYS	18	14
	AT1, GLY, LYS	25	17
10	AT1, GLY, VAL	27	17
11	AT1, ILE, LYS	19	13
12	AT1, LEU, LYS	28	22
13	AT2, AT3, ILE	9	8
14	AT2, AT3, LEU	4	4
	AT2, AT3, LEU	10	9
15	AT2, AT3, VAL	6	6
16	AT2, GLY, LYS	6	6
17	AT3, ILE, VAL	7	5
18	AT3, LEU, LEU	8	7
19	AT3, LEU, TYR	3	3
20	AT3, LEU, VAL	1	1

Table 4.51. Structural motifs found in the TTP dataset

Structural Motif ID	Vertex Types	Frequencies	Distribution (# of coverage PDB)
1	ARG, AT1, AT2	3	2
2	ARG, AT1, GLU	2	1
3	ARG, AT1, GLY	2	1
4	ASP, AT1, AT2	7	6
5	AT1, AT2, GLU	2	2
6	AT1, AT2, GLY	7	5
7	AT1, AT2, LYS	12	10
	AT1, AT2, LYS	10	10
8	AT1, GLY, ILE	2	2
9	AT1, GLY, LYS	4	4
	AT1, GLY, LYS	6	5
	AT1, GLY, LYS	6	5
10	AT1, GLY, VAL	2	2
11	AT1, ILE, LYS	6	5
12	AT1, LEU, LYS	9	6
13	AT2, AT3, ILE	9	8
14	AT2, AT3, LEU	2	2
	AT2, AT3, LEU	4	4
15	AT2, AT3, VAL	2	2
16	AT2, GLY, LYS	1	1
17	AT3, ILE, VAL	0	0
18	AT3, LEU, LEU	3	3
19	AT3, LEU, TYR	0	0
20	AT3, LEU, VAL	3	2

4.6.4. Statistical Significance of the Motifs

We performed bootstrapping to evaluate the statistical significance of the discovered motifs. During the bootstrapping, all amino acids in the protein-Ligand interfaces were put into a bag. Then each amino acid vertex was re-assigned an identity by randomly taking one amino acid from the bag. Then the amino acid was put back into the bag. This process continued until all the amino acid vertices were re-assigned. The identities of ATP subgroup vertices were re-assigned using similar method. After the re-assignment of vertices, the protein-Ligand interfaces were scanned to count how many times each of the motifs occurred. Bootstrapping was performed on the all four datasets separately, repeating 1000 times for each dataset.

Tables 4.52-4.55 (columns 4 and 5) show the average frequency and standard deviation of the bootstrapping for each dataset. The average frequency indicated how many times a motif was expected to occur in the dataset of protein-Ligand interfaces if the vertex identities were randomly assigned. For each dataset, we performed a one-sample t-test to determine whether the observed frequency in the real dataset could have come from a random process mimicked by the bootstrapping. Most of results show that the observed frequencies were not a result of random process. All the t-tests has p value lower than 0.0001 in each of the dataset.

Table 4.52. Statistical significance of the motifs in ATP dataset

Structural Motif ID	Vertex Types	Observed frequency in the real Protein-Ligand interfaces	Average frequency in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	ARG, AT1, AT2	9	1.074	1.123	<0.0001
2	ARG, AT1, GLU	6	0.643	0.895	<0.0001
3	ARG, AT1, GLY	10	1.529	1.491	<0.0001
4	ASP, AT1, AT2	7	0.567	0.793	<0.0001
5	AT1, AT2, GLU	8	0.696	0.877	<0.0001
6	AT1, AT2, GLY	14	1.865	1.579	<0.0001
7	AT1, AT2, LYS	10	1.057	1.103	<0.0001
	AT1, AT2, LYS	11	1.010	1.106	<0.0001
8	AT1, GLY, ILE	7	1.93	1.629	<0.0001
9	AT1, GLY, LYS	9	1.43	1.393	<0.0001
	AT1, GLY, LYS	9	0.937	1.077	<0.0001
	AT1, GLY, LYS	12	1.751	1.573	<0.0001
10	AT1, GLY, VAL	11	2.026	1.690	<0.0001
11	AT1, ILE, LYS	8	1.439	1.412	<0.0001
12	AT1, LEU, LYS	9	1.679	1.584	<0.0001
13	AT2, AT3, ILE	7	1.357	1.266	<0.0001
14	AT2, AT3, LEU	10	2.146	1.644	<0.0001
	AT2, AT3, LEU	6	0.983	1.091	<0.0001
15	AT2, AT3, VAL	7	1.789	1.510	<0.0001
16	AT2, GLY, LYS	12	1.125	1.205	<0.0001
17	AT3, ILE, VAL	8	1.598	1.556	<0.0001
18	AT3, LEU, LEU	8	1.861	1.669	<0.0001
19	AT3, LEU, TYR	6	0.672	0.930	<0.0001
20	AT3, LEU, VAL	9	1.701	1.752	<0.0001

Table 4.53. Statistical significance of the motifs in CTP dataset

Structural Motif ID	Vertex Types	Observed frequency in the real Protein-Ligand interfaces	Average frequency in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	ARG, AT1, AT2	7	2.115	1.475	<0.0001
2	ARG, AT1, GLU	7	1.271	1.261	<0.0001
3	ARG, AT1, GLY	7	2.311	1.705	<0.0001
4	ASP, AT1, AT2	12	1.625	1.385	<0.0001
5	AT1, AT2, GLU	4	1.431	1.277	<0.0001
6	AT1, AT2, GLY	16	3.378	2.095	<0.0001
7	AT1, AT2, LYS	14	1.93	1.485	<0.0001
	AT1, AT2, LYS	17	1.518	1.286	<0.0001
8	AT1, GLY, ILE	3	2.373	1.784	<0.0001
9	AT1, GLY, LYS	2	1.685	1.539	<0.0001
	AT1, GLY, LYS	5	1.092	1.199	<0.0001
	AT1, GLY, LYS	7	2.268	1.724	<0.0001
10	AT1, GLY, VAL	11	3.868	2.331	<0.0001
11	AT1, ILE, LYS	9	1.675	1.425	<0.0001
12	AT1, LEU, LYS	5	2.25	1.725	<0.0001
13	AT2, AT3, ILE	8	2.332	1.662	<0.0001
14	AT2, AT3, LEU	6	3.764	2.200	<0.0001
	AT2, AT3, LEU	2	1.636	1.333	<0.0001
15	AT2, AT3, VAL	6	3.76	2.131	<0.0001
16	AT2, GLY, LYS	4	1.483	1.401	<0.0001
17	AT3, ILE, VAL	1	2.343	1.752	<0.0001
18	AT3, LEU, LEU	2	2.873	2.009	<0.0001
19	AT3, LEU, TYR	1	1.8	1.508	<0.0001
20	AT3, LEU, VAL	0	2.518	1.859	<0.0001

Table 4.54. Statistical significance of the motifs in GTP dataset

Structural Motif ID	Vertex Types	Observed frequency in the real Protein-Ligand interfaces	Average frequency in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	ARG, AT1, AT2	20	3.036	1.838	<0.0001
2	ARG, AT1, GLU	10	2.227	1.639	<0.0001
3	ARG, AT1, GLY	7	4.701	2.468	<0.0001
4	ASP, AT1, AT2	10	2.591	1.733	<0.0001
5	AT1, AT2, GLU	7	2.315	1.661	<0.0001
6	AT1, AT2, GLY	44	6.473	2.916	<0.0001
7	AT1, AT2, LYS	31	3.858	2.167	<0.0001
	AT1, AT2, LYS	34	2.994	1.834	<0.0001
8	AT1, GLY, ILE	16	5.668	2.703	<0.0001
9	AT1, GLY, LYS	8	4.741	2.433	<0.0001
	AT1, GLY, LYS	18	3.124	2.038	<0.0001
	AT1, GLY, LYS	25	6.35	3.016	<0.0001
10	AT1, GLY, VAL	27	7.263	3.261	<0.0001
11	AT1, ILE, LYS	19	3.949	2.214	<0.0001
12	AT1, LEU, LYS	28	4.423	2.360	<0.0001
13	AT2, AT3, ILE	9	3.733	2.077	<0.0001
14	AT2, AT3, LEU	4	5.533	2.582	<0.0001
	AT2, AT3, LEU	10	2.954	1.785	<0.0001
15	AT2, AT3, VAL	6	4.422	2.612	<0.0001
16	AT2, GLY, LYS	6	4.034	2.272	<0.0001
17	AT3, ILE, VAL	7	3.905	2.192	<0.0001
18	AT3, LEU, LEU	8	4.293	2.424	<0.0001
19	AT3, LEU, TYR	3	2.129	1.632	<0.0001
20	AT3, LEU, VAL	1	4.279	2.385	<0.0001

Table 4.55. Statistical significance of the motifs in TTP dataset

Structural Motif ID	Vertex Types	Observed frequency in the real Protein-Ligand interfaces	Average frequency in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
1	ARG, AT1, AT2	3	1.359	1.280	<0.0001
2	ARG, AT1, GLU	2	0.872	1.010	<0.0001
3	ARG, AT1, GLY	2	0.593	1.460	<0.0001
4	ASP, AT1, AT2	7	0.666	0.862	<0.0001
5	AT1, AT2, GLU	2	0.699	0.918	<0.0001
6	AT1, AT2, GLY	7	2.051	1.543	<0.0001
7	AT1, AT2, LYS	12	0.97	1.089	<0.0001
	AT1, AT2, LYS	10	0.972	1.067	<0.0001
8	AT1, GLY, ILE	2	1.584	1.505	<0.0001
9	AT1, GLY, LYS	4	0.971	1.078	<0.0001
	AT1, GLY, LYS	6	0.738	0.971	<0.0001
	AT1, GLY, LYS	6	1.612	1.505	<0.0001
10	AT1, GLY, VAL	2	1.782	1.534	<0.0001
11	AT1, ILE, LYS	6	1.061	1.153	<0.0001
12	AT1, LEU, LYS	9	1.268	1.305	<0.0001
13	AT2, AT3, ILE	9	1.368	1.251	<0.0001
14	AT2, AT3, LEU	2	2.043	1.613	0.3994
	AT2, AT3, LEU	4	0.964	1.051	<0.0001
15	AT2, AT3, VAL	2	1.308	1.265	<0.0001
16	AT2, GLY, LYS	1	0.921	1.060	0.0186
17	AT3, ILE, VAL	0	1.034	1.187	<0.0001
18	AT3, LEU, LEU	3	1.653	1.471	<0.0001
19	AT3, LEU, TYR	0	1.377	1.335	<0.0001
20	AT3, LEU, VAL	3	1.103	1.207	<0.0001

Table 4.52 shows that all motifs occurred in the real protein-ATP interfaces with higher frequencies than the result of a random process, and the differences are statistically significant. Tables 4.53, 4.54, and 4.55 show that the same trend were observed in the protein-CTP, protein-GTP, and protein TTP interfaces for all motifs except motifs 17 and 19. Tables 4.56 and 4.57 show the occurrence of motifs 17 and 19 in the four types protein-ligand interfaces. In the protein-CTP and protein-TTP datasets, motifs 17 and 19 occur in the protein-ligand interfaces with frequencies lower than expected by a random process. In contrast, they both occur in the protein-ATP and protein-GTP interfaces with frequencies higher than expected by a random process. Combining together, these results indicate that motifs 17 and 19 are abundant in protein-ATP and protein-GTP interfaces but are depleted in protein-CTP and protein-TTP interfaces. Since ATP and GTP both have a purine base, and CTP and TTP have a pyrimidine base, these results suggest that motifs 17 and 19 facilitate the interactions between purines and amino acids but are not used in the interactions between pyrimidine and amino acids.

Table 4.56. Statistical significance of motif 17 in four datasets

Dataset	Observed frequency in the real Protein-Ligand interfaces	Average frequency in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
protein-ATP	8	1.598	1.556	<0.0001
protein-CTP	1	2.343	1.752	<0.0001
protein-GTP	7	3.905	2.192	<0.0001
protein-TTP	0	1.034	1.187	<0.0001

Table 4.57. Statistical significance of motif 19 in four datasets

Dataset	Observed frequency in the real Protein-Ligand interfaces	Average frequency in bootstrapping	Standard deviation of bootstrapping	p-value of one-sample t-test
protein-ATP	6	0.672	0.930	<0.0001
protein-CTP	1	1.8	1.508	<0.0001
protein-GTP	3	2.129	1.632	<0.0001
protein-TTP	0	1.377	1.335	<0.0001

5. CONCLUSION

Biologists have strong desire for the ability to engineering proteins that can specifically bind to another structure such as DNA, protein or ligand, which inspires them to look for "recognition codes" that govern the interactions between proteins and another molecular.

Most sequence-based and PWM methods only capture sequence features on the protein interfaces and ignore the crucial spatial attributes of the features. For example in the protein-DNA studies, the pairwise contacting preferences method only takes pair-wise distances between amino acids and nucleotide bases into consideration, which is not enough. Some methods take into consideration the 3-dimensional structures of the interacting molecules, but only consider them separately.

In this work, we implemented an innovative graph representation method to encode protein complex structure. For the example of protein-DNA studies, our method applied the 3D distance and relative position among amino acids and nucleotides to the construction of graph. Then we discovered structural motifs that were favored in the protein-DNA, protein-protein and protein-ligand interactions. Such motifs include more information than the traditional amino acid-base contacting pairs. Thus, they can provide more accurate prediction on protein-DNA, protein-protein and protein-ligand recognition. The biological and statistical significance of the motifs were confirmed using evolutionary conservation analysis and bootstrapping. We also performed many other tests to evaluate our motifs' critical roles in the interactions. For example, we compared our motifs with experimentally verified hotspots. We also compared our method with other computational prediction method to assess the effectiveness of the method.

Our results confirmed that the graph motifs discovered in this study play important roles in protein-DNA, protein-protein and protein-ligand interactions. Using product graph, we transformed the structural motif discovery problem into a search for maximal cliques. This study

sheds light on the recognition codes from a new angle by representing the recognition codes in form of structural motifs that span the protein-DNA, protein-protein and protein-ligand interfaces. We believe that the proposed graph method will be a very helpful tool for studying the protein complexes interaction and other type of molecular interactions.

REFERENCES

[1] Jones, S.; van Heyningen, P.; Berman, H.M.; Thornton, J.M. protein–DNA interactions: A structural analysis. J. Mol. Biol. 1999, 287, 877–896.

[2] Jones, S.; Barker, J.A.; Nobeli, I.; Thornton, J.M. Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res. 2003, 31, 2811–2823.

[3] Kono, H.; Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins 1999, 35, 114–131.

[4] Luscombe, N.M.; Laskowski, R.A.; Thornton, J.M. Amino acid-base interactions: A three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res. 2001, 29, 2860–2874.

[5] Mandel-Gutfreund, Y.; Margalit, H. Quantitative parameters for amino acid-base interaction: Implications for prediction of protein–DNA binding sites. Nucleic Acids Res. 1998, 26, 2306–2312.

[6] Olson, W.K.; Gorin, A.A.; Lu, X.J.; Hock, L.M.; Zhurkin, V.B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. Proc. Natl. Acad. Sci. USA 1998, 95, 11163–11168.

[7] Orengo, C.A.; Michie, A.D.; Jones, S.; Jones, D.T.; Swindells, M.B.; Thornton, J.M. CATH—A hierarchic classification of protein domain structures. Structure 1997, 5, 1093–1108.

[8] Ponting, C.P.; Schultz, J.; Milpetz, F.; Bork, P. SMART: Identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res. 1999, 27, 229–232.

[9] Hwang, S.; Gou, Z.; Kuznetsov, I.B. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 2007, 23, 634–636.

[10] Yan, C.; Terribilini, M.; Wu, F.; Jernigan, R.L.; Dobbs, D.; Honavar, V. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinform. 2006, 7, doi:10.1186/1471-2105-7-262.

[11] Ahmad, S.; Sarai, A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinform. 2005, 6, doi:10.1186/1471-2105-6-33.

[12] Wu, J.; Liu, H.; Duan, X.; Ding, Y.; Wu, H.; Bai, Y.; Sun, X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 2009, 25, 30–35.

[13] Carson, M.B.; Langlois, R.; Lu, H. NAPS: A residue-level nucleic acid-binding prediction server. Nucleic Acids Res. 2010, 38, W431–W435.

[14] Alibes, A.; Serrano, L.; Nadra, A.D. Structure-based DNA-binding prediction and design. Methods Mol. Biol. 2010, 649, 77–88.

[15] Li, B.Q.; Feng, K.Y.; Ding, J.; Cai, Y.D. Predicting DNA-binding sites of proteins based on sequential and 3D structural information. Mol. Genet. Genomics 2014, 289, 489–499.

[16] Li, T.; Li, Q.Z.; Liu, S.; Fan, G.L.; Zuo, Y.C.; Peng, Y. PreDNA: Accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics 2013, 29, 678–685.

[17] Xiong, Y.; Xia, J.; Zhang, W.; Liu, J. Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. PLoS One 2011, 6, e28440.

[18] Dominguez, C.; Boelens, R.; Bonvin, A.M. HADDOCK: A protein–protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. 2003, 125, 1731–1737. [CrossRef] [PubMed]

[19] Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A.A.; Aflalo, C.; Vakser, I.A. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA 1992, 89, 2195–2199. [CrossRef] [PubMed]

[20] Ritchie, D.W.; Kemp, G.J. Protein docking using spherical polar Fourier correlations. Proteins 2000, 39, 178–194. [CrossRef]

[21] Gabb, H.A.; Jackson, R.M.; Sternberg, M.J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. 1997, 272, 106–120. [CrossRef] [PubMed]

[22] Tuszynska, I.; Bujnicki, J.M. DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking. BMC Bioinform. 2011, 12. [CrossRef] [PubMed]

[23] Pabo, C.O. and Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J. Mol. Biol., 301, 597–624.

[24] Suzuki, M., Gerstein, M. and Yagi, N. (1994) Stereochemical basis of DNA recognition by Zn fingers. Nucleic Acids Res., 22, 3397–3405.

[25] Kortemme, T., Morozov, A.V. and Baker, D. (2003) An orientation dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J. Mol. Biol., 326, 1239–1259.

[26] Siggers, T.W. and Honig, B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Nucleic Acids Res., 35, 1085–1097.

[27] Siggers, T.W., Silkov, A. and Honig, B. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. J. Mol. Biol., 345, 1027–1045.

[28] Morozov, A.V.; Havranek, J.J.; Baker, D.; Siggia, E.D. protein–DNA binding specificity predictions with structural models. Nucleic Acids Res. 2005, 33, 5781–5798.

[29] Szilagyi, A.; Skolnick, J. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J. Mol. Biol. 2006, 358, 922–933.

[30] Gao, M.; Skolnick, J. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput. Biol. 2009, 5, e1000567.

[31] Si, J.; Zhang, Z.; Lin, B.; Schroeder, M.; Huang, B. MetaDBSite: A meta approach to improve protein DNA-binding sites prediction. BMC Syst. Biol. 2011, 5, doi:10.1186/1752-0509-5-S1-S7.

[32] Zhou, Q.; Liu, J.S. Extracting sequence features to predict protein–DNA interactions: A comparative study. Nucleic Acids Res. 2008, 36, 4137–4148.

[33] Rhodes, Daniela, et al. "Towards an understanding of protein-DNA recognition." Philosophical Transactions of the Royal Society of London B: Biological Sciences 351.1339 (1996): 501-509.

[34] Hall, Traci M. Tanaka. "Multiple modes of RNA recognition by zinc finger proteins." Current opinion in structural biology 15.3 (2005): 367-373.

[35] Choo, Yen, and Aaron Klug. "Physical basis of a protein-DNA recognition code." Current opinion in structural biology 7.1 (1997): 117-125.

[36] Miller, Jeffrey C., and Carl O. Pabo. "Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition." Journal of molecular biology 313.2 (2001): 309-315.

[37] Liu, Limin Angela, and Philip Bradley. "Atomistic modeling of protein–DNA interaction specificity: progress and applications." Current opinion in structural biology 22.4 (2012): 397-405.

[38] Baldwin, E.P., Martin, S.S., Abel, J., Gelato, K.A., Kim, H., Schultz, P.G. and Santoro, S.W. (2003) A specificity switch in selected cre recombinase variants is mediated by macromolecular plasticity and water. Chem. Biol., 10, 1085–1094.

[39] Stormo, G.D. (2000) DNA binding sites: representation and discovery. Bioinformatics, 16, 16–23.

[40] Staden, R. (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res., 12, 505–519.

[41] Siggers, Trevor, and Raluca Gordân. "Protein–DNA binding: complexities and multi-protein codes." Nucleic acids research (2013): gkt1112.

[42] Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D. and Benos, P.V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res., 33, W389–W392.

[43] Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W. 3rd and Bulyk, M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol., 24, 1429–1435.

[44] Siggers, T., Chang, A.B., Teixeira, A., Wong, D., Williams, K.J., Ahmed, B., Ragoussis, J., Udalova, I.A., Smale, S.T. and Bulyk, M.L. (2012) Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-kappaB family DNA binding. Nat. Immunol., 13, 95–102.

[45] Nakagawa, S., Gisselbrecht, S.S., Rogers, J.M., Hartl, D.L. and Bulyk, M.L. (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc. Natl Acad. Sci. USA, 110, 12349–12354.

[46] Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C.D., Gossett, A.J., Hasinoff, M.J., Warren, C.L. et al. (2008) A library of yeast transcription factor motifs reveal a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol. Cell, 32, 878–887.

[47] Wolberger C, Vershon AK, Lui B, Johnson AD, Pabo CO: Crystal structure of a MATa2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell 1991, 67:517-528.

[48] Suzuki M: Common features in DNA recognition helices of eukaryotic transcription factors. EMBO J 1993, 12:3221-3226.

[49] Shih, W.; Chai, S. Data-Driven vs. Hypothesis-Driven Research: Making sense of big data. In Academy of Management Proceedings (Vol. 2016, No. 1, p. 14843). Briarcliff Manor, NY 10510: Academy of Management.

[50] Anderson, C. The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07. 2008.

[51] Mazzocchi, F. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO reports, 16(10), 1250-1255. 2015.

[52] Van Helden, P. Data-driven hypotheses. EMBO reports, 14(2), 104-104. 2013.

[53] Kraus, W. L. Would You Like A Hypothesis with Those Data? Omics and the Age of Discovery Science. 2015.

[54] Evans, J.; Rzhetsky, A. Machine science. Science, 329(5990), 399-400. 2010.

[55] Berman, H.M., et al., The Protein Data Bank. Nucl Acids Res, 2000. 28(1): p. 235-242.

[56] Wang, G. and R.L.J. Dunbrack, PISCES: a protein sequence culling server. Bioinformatics, 2003. 19: p. 1589-1591.

[57] Van Dijk, M. and A.M.J.J. Bonvin, A protein–DNA docking benchmark. Nucleic Acids Research, 2008. 36(14): p. e88.

[58] Jones, S., et al., Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucl Acids Res, 2003. 31(24): p. 7189-7198.

[59] Hubbard, S.J., NACCESS. 1993, Department of Biochemistry and Molecular Biology, University College, London.

[60] Cordella, L., et al., A (sub) graph isomorphism algorithm for matching large graphs. IEEE Trans Pattern Anal Mach Intell, 2004. 26(10): p. 1367-1372.

[61] Hall, M., et al., The WEKA Data Mining Software: An Update. SIGKDD Explorations, 2009. 1(1).

[62] Bairoch, A., et al., The Universal Protein Resource (UniProt). Nucl. Acids Res., 2005. 33: p. D154-D159.

[63] Gabb, H.A., R.M. Jackson, and M.J.E. Sternberg, Modelling protein docking using shape complementarity, electrostatics and biochemical information1. Journal of Molecular Biology, 1997. 272(1): p. 106- 120.

[64] Parisien, M., K. Freed, and T. Sosnick, On Docking, Scoring and Assessing Protein-DNA Complexes in a Rigid-Body Framework. PLoS ONE, 2012. 7(2): p. e32647.